روش‌های تحلیل رگرسیونی برای داده‌های بعد بالا

منیره معنوی[۱]، مهدی روزبه[۲]

تاریخ دریافت: ۹۹/۴/۳۱

تاریخ پذیرش: ۹۹/۱۱/۱

چکیده:

با پیشرفت علم، دانش و تکنولوژی، روش‌های جدید و جامع برای اندازه‌گیری، جمع‌آوری و ثبت اطلاعات ابداع‌شده‌اند، که منجر به ظهور و توسعه داده‌های بعد بالا شده‌اند. مجموعه داده‌های بعد بالا، یعنی مجموعه داده‌هایی که در آن تعداد متغیرهای توضیحی بسیار بزرگ‌تر از تعداد مشاهدات است، به‌سادگی و با روش‌های سنتی و کلاسیک، مانند روش کمترین توان‌های دوم معمولی، نمی‌توانند تحلیل شوند و تفسیرپذیری آن امری بسیار پیچیده خواهد بود. اگرچه درصورتی‌که فرضیات اساسی برقرار باشند، برآورد کمترین توان‌های دوم معمولی بهترین روش برآورد در تحلیل رگرسیونی است ولی برای داده‌های بعد بالا قابل‌استفاده نبوده و در این شرایط مستلزم به‌کارگیری روش‌هایی نوینی هستیم. در این مقاله در ابتدا، به مشکلات روش‌های کلاسیک در تحلیل داده‌های بعد بالا اشاره می‌شود و سپس، به معرفی و توضیح روش‌های تحلیل رگرسیونی متداول و امروزی مانند روش‌های تحلیل مؤلفه اصلی و تاوانیده برای داده‌های بعد بالا پرداخته می‌شود. در انتها یک مطالعه شبیه‌سازی و تحلیل داده واقعی برای بررسی و مقایسه روش‌های اشاره‌شده در داده‌های بعد بالا انجام می‌گردد.

واژه‌های کلیدی: تحلیل مؤلفه‌های اصلی، مجموعه داده‌های بعد بالا، روش کمترین توان‌های دوم تاوانیده.

۱ مقدمه

امروزه با گسترش روزافزون علم، دانش و فناوری، روش‌های نوین و دقیقی برای اندازه‌گیری، جمع‌آوری و ثبت اطلاعات ابداع‌شده و این امر باعث ظهور و گسترش داده‌های بعد بالا شده است. از اواخر سال ۱۹۹۰ کار با این مجموعه داده‌ها یعنی داده‌هایی که در آن تعداد متغیرهای توضیحی (p) بسیار بیشتر از تعداد مشاهدات (n) است، شروع شد [۱۰]. افزایش بیش‌ازپیش ابعاد داده‌ها به یک مسئله اساسی در مبحث داده‌کاوی تبدیل‌شده است. تحلیل و تفسیر این داده‌ها امری دشوار و بسیار قابل‌تأمل است به‌طوری‌که واسرمن [۲۵] در کتاب خود به این امر اشاره نموده و از این پدیده تحت عنوان نفرین ابعاد[۳] یادکرده است. اصطلاحی که معمولاً به بلمن [۶] نسبت داده می‌شود. این بدان معنی است که با افزایش بعد متغیرها، برآورد پارامترهای مربوطه بسیار دشوار می‌شود. حداقل دو نوع از این نفرین وجود دارد. اولین مورد، نفرین محاسباتی ابعاد است. این امر به این واقعیت اشاره دارد که بار محاسباتی برخی از روش‌ها می‌تواند به‌صورت نمایی با بعد افزایش یابد. دومین مورد که واسرمن آن را نفرین آماری ابعاد نامید به این صورت است که اگر داده‌ها دارای بعد d باشند، پس نمونه‌ای با اندازه n موردنیاز است که به‌صورت نمایی با d رشد کند، تا بتواند قابل‌تحلیل و تفسیر با روش‌های کلاسیک باشد.

در گذشته، کار با داده‌های کلاسیک که دارای حداکثر چند ده متغیر توضیحی بوده‌اند، بسیار ساده بود و روش‌های کلاسیک نظیر کمترین توان‌های دوم نتایج قابل قبولی ارائه می‌دادند. اما این روش‌ها در حضور داده‌های بعد بالا برآوردی ارائه نمی‌دهند. در تحلیل رگرسیونی داده‌های بعد بالا تعداد زیاد متغیرهای توضیحی، محقق را با چالشی جدی روبه‌رو می‌کند و همیشه جدالی بین دقت، سرعت و هزینه وجود دارد. از سوی دیگر، وجود متغیرهایی در مدل که با متغیر پاسخ ارتباطی ندارند، منجر به بیش‌برازشی[۴] می‌شود. چنین مدلی، مدلی بسیار پیچیده برای داده‌ها است. در این شرایط، مدل با تغییرات جهشی سعی در پوشش داده‌های حاصل از نمونه و حتی مقدارهای خطاها می‌کند. درحالی‌که این مدل باید منعکس‌کننده‌ی رفتار جامعه باشد. به‌کارگیری تمامی متغیرهای توضیحی زمانی طولانی صرف کرده و هزینه‌های محاسباتی بسیاری را بر محقق تحمیل می‌کند. در این شرایط این ابهام مطرح می‌شود که شاید حضور تمامی متغیرها برای برازش مدل الزامی نباشد و بدین‌سان اغلب محققان با در نظر گرفتن فرض تنکی[۵] سعی در کاهش بعد دارند و از زیرمجموعه‌ای مناسب از متغیرها برای برازش

[۱] دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.

[۲] هیئت‌علمی گروه آمار، دانشگاه سمنان، سمنان، ایران mahdi.roozbeh@semnan.ac.ir.

[3] Curse of Dimensionality

[4] Overfiting

[5] Sparsity

مدل کمک می‌گیرند و ازاین‌رو علاقه‌مند به شناسایی این زیرمجموعه مناسب هستند [۷]. حتی امروزه مدل‌های رگرسیونی لوژستیک نیز در این زمینه مورد اهمیت ویژه‌ای قرارگرفته‌اند [۲۴].

مثال ۱.۱. در مطالعات ژنتیکی که در اکثر آن‌ها ارتباط بین تغییرات *DNA* و بیماری هدف موردمطالعه می‌باشد، ۹۹٫۹ درصد از سکانس *DNA* انسان‌ها با یکدیگر یکسان است و بیش از ۸۰درصد از آن ۰٫۱ درصد باقی‌مانده به *SNP* (پلی‌مرفیسم تک نوکلئوتیدی[6]) مرتبط است. میلیون‌ها *SNP* با هم تعامل دارند تا فنوتیپ نهایی بیماری را تعیین کنند. این تعداد زیاد *SNP* ها و امکان وجود اثرهای متقابل بین عوامل ژنتیکی با هم و با عوامل محیطی یک چالش را در تحلیل این‌که کدام یک از این *SNP* ها در ارتباط با بیماری هستند، پدید آورده است. هدف اصلی در تحلیل این اطلاعات تعیین این‌که کدام *SNP* یا مجموعه‌ای از آن‌ها روی بیماری مؤثر هستند، نیست چراکه بسیاری از *SNP* ها اثرهای حاشیه‌ای ناچیزی دارند ولی اثرات متقابل آن‌ها بسیار قوی است. هدف اصلی اولویت‌بندی *SNP* ها بر اساس اهمیت و نقش آن‌ها در ارتباط با بیماری به‌منظور مطالعه و بررسی بیشتر آن‌هاست. در این شرایط برای رسیدن به این هدف در نظر گرفتن فرض تنکی الزامی است [۴].

در ادامه به معرفی مدل رگرسیونی بپردازیم. مدل رگرسیونی خطی ساده به‌صورت

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \qquad (۱)$$

است، به‌طوری‌که در آن $\mathbf{y} = (y_۱, \ldots, y_n)^\top$، بردار متغیر پاسخ، $\mathbf{X} = (\mathbf{x}_۱, \ldots, \mathbf{x}_p)_{n\times p}$ ماتریس مشاهدات متغیرهای توضیحی با $\mathbf{x}_i = (x_{۱i}, \ldots, x_{ni})^\top$، $\beta = (\beta_۱, \ldots, \beta_p)^\top$ بردار پارامترها و $\varepsilon = (\varepsilon_۱, \ldots, \varepsilon_n)^\top$ بردار خطا است.

۲ روش‌های تحلیل رگرسیونی برای داده‌های بعد بالا

در این بخش، چندین روش برای رویارویی با داده‌های بعد بالا معرفی می‌شود. برخی از این روش‌ها فرض تنکی را در نظر می‌گیرند و برخی در برنمی‌گیرند.

۱.۲ روش‌های کلاسیک

رگرسیون بهترین زیرمجموعه[7] یکی از روش‌های کلاسیک است. ایده اصلی این روش، ساخت مدلی با زیرمجموعه‌ای از متغیرها است. در اصل این زیرمجموعه، کوچک‌ترین زیرمجموعه‌ای از متغیرها است که به‌اندازه کافی تغییرات متغیر پاسخ را توجیه می‌کند. برای تعیین مدل با این روش ابتدا مدلی که تنها حاوی عرض از مبدأ بوده در نظر گرفته می‌شود. سپس تمامی زیرمجموعه‌های تک عضوی و دو عضوی، سه عضوی و ... در نظر گرفته می‌شود. و پس‌ازآن مدل‌های رگرسیونی ایجادشده و برازش داده می‌شوند. در قدم آخر تمامی مدل‌های ایجادشده با یکی از معیارهای سنجش نیکویی برازش مدل نظیر ضریب تعیین، تابع زیان، مجموع توان‌های دوم خطا، C_p مالوس[8]، اعتبارسنجی متقابل[9] و ... مقایسه می‌شوند و بهترین آن‌ها به‌عنوان مدل نهایی انتخاب می‌شود.

در این قسمت سؤال بسیار مهمی پیش می‌آید: اندازه زیرمجموعه‌ها در چه بازه‌ای تغییر می‌کند؟ این سؤال به‌سادگی از طریق رابطه

$$\min\{۰, ۱, \ldots, M\}, \quad M \le \min\{n-۱, (p+۱)\}$$

پاسخ داده می‌شود. که در آن، M اندازه زیرمجموعه‌ها، n تعداد مشاهدات و p تعداد متغیرهای توضیحی است.

رگرسیون زیرمجموعه ضرایب متغیرهایی که در زیرمجموعه انتخابی وجود نداشته باشند، صفر در نظر می‌گیرد و یا به‌عبارت‌دیگر آن را متورم می‌کند [۹]. این روش، به دو دلیل کاهش واریانس و سادگی تفسیر نتایج (برای تعداد متغیرهای توضیحی کم) موردتوجه قرارگرفته است [۹]. رقیب اول رگرسیون زیرمجموعه از نظر کاهش واریانس، روش ستیغی [۱۷] است. رگرسیون زیرمجموعه می‌تواند دقت پیش‌گویی را افزایش دهد اما اگر تعداد متغیرهای موجود در مدل، M، بزرگ باشد، این مسئله ابهام‌برانگیز است.

یک جنبه منفی رگرسیون بهترین زیرمجموعه بی‌ثباتی آن در اثر انحراف‌های کوچک در داده‌هاست. به‌عنوان‌مثال، اگر تنها مشاهده $(y_n, \mathbf{x}_n)$ از مجموعه داده‌ها حذف شوند و از روش‌های انتخاب یکسان استفاده شود دو زیرمجموعه نهایی که با استفاده از این مشاهده و بدون استفاده از آن تعیین‌شده‌اند، بسیار متفاوت خواهند بود. بنابراین منجر به تغییر شدید در معادله پیش‌گویی خواهد شد. این در حالی است که اگر از روش ستیغی استفاده شود (البته با پارامتر ستیغی یکسان) برآوردگرهای حاصل از حذف مشاهده $(y_n, \mathbf{x}_n)$ و برآوردگر حاصل بدون حذف آن تفاوت چندانی نخواهند داشت [۹].

یکی دیگر از نقاط ضعف این روش زمانی است که تعداد متغیرهای توضیحی بیشتر از ۴۰ باشد. در این شرایط به دلیل پیچیدگی ترکیباتی، رگرسیون بهترین زیرمجموعه به یک مسئله دشوار و پیچیده تبدیل می‌شود،

[6] Single-nucleotide polymorphism
[7] Best-subset regression
[8] C_p-Mallows
[9] Cross validation

زیرا 2^p زیرمجموعه وجود دارد که حل آن هزینه‌های محاسباتی زیادی را به همراه دارد. حتی اگر محاسبات نیز انجام‌پذیر باشد، با چنین فضای جست‌وجوی عظیمی، واریانس این روش بسیار بالا بوده و این امر بدین معناست که این روش تنها برای p های نسبتاً کوچک کاربردی است.

روش دیگری که در این بخش مطرح می‌شود، رگرسیون گام‌به‌گام [10] نام دارد. رگرسیون گام‌به‌گام به سه نوع پس‌رو [11]، پیشرو [12] و مرحله‌ای [13] (ترکیبی از هر دو روش پیشرو و پس‌رو) تقسیم می‌شود.

روش پس‌رو، مدلی را که در آن تمامی متغیرها حضور دارند در نظر گرفته و با ملاک قرار دادن یکی از معیارهای سنجش اقدام به بررسی برای حذف متغیرهای بی‌تأثیر در مدل می‌نماید.

روش پیشرو، مدلی را که تنها در آن عرض از مبدأ وجود دارد در نظر گرفته و با ملاک قرار دادن یکی از معیارهای سنجش اقدام به یافتن تأثیرگذارترین متغیر می‌نماید و آن را به مدل اضافه می‌کند و این روند را تا زمانی که معیار سنجش معنی‌دار نشود، ادامه می‌دهد.

می‌توان گفت که روش مرحله‌ای ترکیبی از روش‌های پس‌رو و پیشرو می‌باشد که در آن، در هر مرحله هم‌زمان با حذف یک متغیر از مدل بر اساس معیار سنجش، وجود همان متغیر در مدل بر اساس معیار سنجش بررسی می‌شود [۱۵].

روش رگرسیون گام‌به‌گام پیشرو از یک اصلاح ساده در روش رگرسیون بهترین زیرمجموعه به‌دست‌آمده است. در رگرسیون گام‌به‌گام پیشرو برخلاف روش رگرسیون بهترین زیرمجموعه، زیرمجموعه‌ها آشیانه‌ای (تودرتو) هستند و این باعث می‌شود تا روش گام‌به‌گام پیشرو قابل‌کنترل‌تر باشد و برای مدل‌هایی که تعداد متغیرها بیشتر از ۴۰ هست، نیز انجام‌پذیر باشد. اما اگر تعداد متغیرها بسیار زیاد باشد، برای مثال بیش از هزار متغیر توضیحی در مدل حضورداشته باشد، این روش بسیار پیچیده، خسته‌کننده و زمان‌گیر است [۱۰].

امروزه هزینه‌های محاسباتی بسیار زیاد روش‌های کلاسیک، تحلیل رگرسیونی داده‌های بعد بالا را با چالشی بسیار جدی مواجه ساخته است و شیرینی سادگی تفسیرپذیری مدل را به کام محققان تلخ می‌کند. این روش‌ها از منظر محاسباتی اصلاً مقرون‌به‌صرفه نیستند و همان‌طور که ذکر شد، محاسبات این روش‌های ذکرشده عموماً بسیار پیچیده، گیج‌کننده و زمان‌براست. اما مهم‌ترین اشکال این روش‌ها، ناپایداری آن‌ها است زیرا آن‌ها فرآیندی گسسته هستند که با کوچک‌ترین تغییری در داده‌ها نتیجه انتخاب مدل بسیار متفاوت خواهد شد و دقت پیش‌گویی کاهش می‌دهند. بنابراین باید در صحت پیش‌گویی‌ها تردید نمود. برای برطرف کردن این مشکلات، می‌توان از روش‌های دیگری که در ادامه مطرح می‌شود، کمک گرفت [۲].

۲.۲ روش مؤلفه‌های اصلی

روش مؤلفه‌های اصلی [14]، روشی بسیار ساده و کاربردی است که برای نخستین بار توسط پیرسون [۲۰] ریاضی‌دان انگلیسی مطرح شد و در سال ۱۹۳۳ به‌طور مستقل توسط هتلینگ به‌منظور تحلیل ساختارهای ماتریس‌های واریانس کوواریانس و ضریب همبستگی توسعه داده شد [۱۸]. ایده اصلی این روش کاهش ابعاد داده‌ها به نحوی است که حداکثر اطلاعات ممکن موجود در داده‌ها محفوظ بماند. همانند بسیاری از روش‌های چندمتغیره تا قبل از ظهور رایانه و نرم‌افزارهای مرتبط به دلیل پیچیدگی محاسبات به‌صورت گسترده مورداستفاده واقع نشد. اما اکنون تقریباً در تمامی نرم‌افزارهای آماری و ریاضیاتی بسته‌های محاسباتی آن توسط محققان تدوین‌شده است. طبیعتاً پس از پیدایش نرم‌افزارها و بسته‌های رایانه‌ای این روش بسیار محبوب و رایج گردید.

هدف اصلی این روش یافتن ترکیب‌های خطی از متغیرهای توضیحی است که واریانس را به بیشینه خود برساند. این روش بر پایه مقادیر ویژه و بردارهای ویژه ماتریس واریانس-کوواریانس یا ماتریس همبستگی استوار است. روش مؤلفه اصلی، معمولاً یک تجزیه نهایی تلقی نمی‌شود بلکه به‌عنوان ابزاری میانی برای مطالعه و بررسی‌های بیشتر مورداستفاده قرار می‌گیرد.

از دید ریاضیات این روش، یک تبدیل خطی متعامد است که داده‌ها را به دستگاه مختصات جدید می‌برد به‌طوری‌که بزرگ‌ترین واریانس داده‌ها بر روی اولین محور مختصات، دومین بزرگ‌ترین واریانس بر روی دومین محور مختصات قرار می‌گیرد و به همین ترتیب برای سایر متغیرها ادامه می‌یابد. این امر از طریق تبدیل متغیرها به متغیرهای جدیدی رخ می‌دهد که مؤلفه‌های اصلی نامیده می‌شوند. مؤلفه‌های اصلی ناهمبسته بوده و به ترتیبی اولویت‌بندی می‌شوند که تعداد اندکی از آن‌ها اغلب تغییرات موجود در متغیرهای اولیه را با خود به همراه دارند.

از کاربردهای دیگر تحلیل مؤلفه‌های اصلی می‌توان به تبدیل متغیرهای همبسته به متغیرهای ناهمبسته، یافتن ترکیبات خطی با تغییرپذیری نسبی بزرگ یا کوچک، کاهش حجم داده‌ها و تفسیر ساده‌تر آن‌ها اشاره نمود.

گام بعدی پس از ساخت مؤلفه‌ها تعیین تعداد مؤلفه‌هایی است که برای

[10] Stepwise regression
[11] Backward
[12] Forward
[13] Stagewise
[14] Principle component

ایجاد مدل رگرسیونی مورداستفاده می‌گیرد. در واقع در این مرحله کاهش بعد صورت می‌گیرد. انتخاب تعداد مناسب مؤلفه‌ها امری بسیار ضروری و حائز اهمیت است، چراکه هدف هر محقق در این مرحله نگهداری کمترین تعداد مؤلفه‌ها با حفظ بیشترین اطلاعات و تغییرپذیری موجود در مجموعه داده‌ها به‌منظور دسترسی به بالاترین دقت با صرف کمترین زمان ممکن است. روش‌های متعددی برای تعیین تعداد مناسب مؤلفه‌ها وجود دارد که در ادامه به برخی از این موارد به‌اختصار اشاره می‌شود.

یکی از روش‌های تعیین مؤلفه‌ها توجه به تغییرپذیری بیان‌شده توسط مؤلفه‌ها است. در واقع در این روش تعدادی از مؤلفه‌ها که بتوانند درصد قابل‌توجهی از واریانس (تغییرپذیری) کل را توجیه کنند، انتخاب می‌شوند. البته باید توجه داشت که لازم است این درصد نسبت به تعداد مؤلفه‌ها به‌صرفه باشد. برای مثال اگر ۱۰۰ متغیر توضیحی در مجموعه داده‌ها وجود داشته باشد و با اتخاذ ۲۰ مؤلفه ۷۰٪ درصد تغییرات واریانس کل و با اتخاذ ۴۰ مؤلفه ۷۵٪ درصد تغییرات واریانس کل توجیه شود، افزودن ۲۰ مؤلفه دیگر برای دست یافتن تنها به ۵٪ درصد تغییرات بیشتر به‌صرفه نخواهد بود. ولی اگر با در نظر گرفتن ۲۳ مؤلفه ۸۰٪ درصد تغییرات واریانس کل توجیه شود، انتخاب بسیار مناسب و معقولی خواهد بود.

روش دیگر، انتخاب تعداد مؤلفه‌هایی است که واریانس آن‌ها بزرگ‌تر یا مساوی متوسط واریانس کل است. لازم به ذکر است که اگر از ماتریس همبستگی استفاده شود، مؤلفه‌هایی که واریانس آن‌ها بزرگ‌تر یا مساوی ۱ است، انتخاب می‌شوند.

علاوه بر دو روش ذکرشده روش‌های شهودی جالبی نیز وجود دارند. یکی از این روش‌ها استفاده از نمودار بازو[15] است. در این نمودار مقادیر ویژه هر مؤلفه (λ_i) در برابر i رسم می‌شود. با مشخص کردن ناحیه‌ای از نمودار که شیب آن به‌طور ناگهانی کند شده و شکسته می‌شود، تعداد مؤلفه‌ها تعیین می‌شود. البته بعضی محققان لگاریتم λ_i در برابر i رسم می‌کنند. لگاریتم تابعی یک‌به‌یک است که مقیاس را کاهش می‌دهد بنابراین استفاده از این روش تفاوت چندانی با نمودار بازو ندارد، اما استفاده از نمودار بازو سودمندتر خواهد بود چراکه جهش‌های ناگهانی در شیب نمودار بهتر نمایش داده می‌شود و بدین‌سان انتخاب تعداد مؤلفه‌ها ساده‌تر خواهد بود.

لازم به ذکر است که ممکن است هر کدام از روش‌های مطرح‌شده در بالا پاسخ‌های متفاوتی ارائه دهند که بنا به نوع هدف و سلیقه محقق یکی از این روش‌ها برای انتخاب تعداد مناسب متغیرها به کار گرفته می‌شود.

یکی از نکات بسیار ضروری در استفاده از روش مؤلفه‌های اصلی استانداردسازی متغیرهاست. اهمیت این امر بدین‌سان است که متغیرهای توضیحی اولیه ممکن است مقیاس‌های گوناگونی داشته باشند. به‌عنوان‌مثال می‌توان به یک مجموعه داده که شامل متغیرهایی با یکاهای گالون، کیلومتر، سال نوری و ... است، اشاره کرد. بدیهی است که مقدار واریانس این متغیرها اعداد بزرگی خواهد بود. اعمال مؤلفه‌های اصلی روی متغیرهای استاندارد نشده منجر به تأثیر عظیمی بر روی متغیرهایی که واریانس بزرگی دارند، می‌شود و این امر به‌نوبه خود می‌تواند منجر به وابستگی مؤلفه اصلی به متغیرهای دارای واریانس بالا شود که بسیار نامطلوب است. البته برای حل این مشکل می‌توان از ماتریس همبستگی نیز کمک گرفت زیرا ماتریس همبستگی برخلاف ماتریس واریانس-کوواریانس مقیاس پایاست.

۳.۲ روش کمترین توان‌های دوم تاوانیده

برآوردگرهای تاوانیده از به‌کارگیری روش‌های کمترین توان‌های دوم یا ماکسیمم درستنمایی تاوانیده در مدل‌بندی رگرسیونی حاصل می‌شود. در حقیقت این برآوردگرها از روش بهینه‌سازی یک تابع درجه دوم (البته نه همیشه) نسبت به یک برآوردگر تاوانیده به دست می‌آید. ایده اصلی در این برآوردگرها، این است که متغیرهای توضیحی کم‌اهمیت در یک مدل تنک به سمت صفر منقبض شود. در اکثر روش‌های تاوانیده این متغیرها از مدل خارج می‌شوند و همین امر موجب ساده‌تر شدن تفسیرپذیری مدل خواهد شد [۱].

تابع هدف در روش کمترین توان‌های دوم معمولی در مدل رگرسیونی چندگانه (۱) به‌صورت زیر است:

$$\min_{\beta\in\mathbb{R}^p}\left\{(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta)\right\}. \qquad (۲)$$

با محاسبه مشتق تابع هدف عبارت (۲) نسبت به بردار پارامتر β و برابر صفر قرار دادن آن و حل معادله حاصل، برآوردگر کمترین توان‌های دوم به‌سادگی یافت می‌شود. به‌عبارت‌دیگر در این روش خطا برحسب β کمینه می‌شود. واضح است که اگر مقادیر واقعی β بزرگ باشد، برآوردگر حاصل، فارغ از احتساب بزرگی β، فاصله زیادی از مقدار واقعی β خواهد داشت و به همین سبب خطای برآورد زیادشده و دقت آن کاهش می‌یابد [۱]. یک روش بسیار سودمند و تقریباً ساده برای مقابله با این مشکل، تاوان دادن مقادیر بزرگ β است بدین‌سان به‌جای کمینه کردن تابع هدف روش کمترین توان‌های دوم تابع هدف روش کمترین توان‌های دوم تاوانیده یعنی

$$\min_{\beta\in\mathbb{R}^p}\left\{(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta)\right\}, \quad s.t. \quad p(\beta)\le t \qquad (۳)$$

کمینه می‌شود. $t \geq ۰$ پارامتر تنظیم‌کننده[16] نامیده می‌شود که اگر برابر صفر باشد، مدل تنها شامل عرض از مبدأ است و اگر بی‌نهایت باشد، مدل کامل

[15] Scree diagram
[16] Tuning parameter

(با حضور تمامی متغیرها) خواهد بود. $p(\beta)$ نیز تابعی از بردار پارامتر است که تابع تاوان[17] نامیده شده و با توجه به نوع آن روش‌های تاوانیده متفاوتی ایجاد می‌شود. برای یافتن کمینه (بیشینه) یک تابع چند متغیره که با یک یا چند محدودیت مواجه است، روش لاگرانژ به کار گرفته می‌شود و با استفاده از روش لاگرانژ چند متغیره عبارت (۳) به‌صورت زیر بازنویسی می‌شود [۱۱]:

$$\min_{\beta\in\mathbb{R}^p}\Big\{(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta)+\lambda p(\beta)\Big\}. \quad (۴)$$

در عبارت (۴) پارامتر نامنفی λ، پارامتر منظم‌سازی[18] یا پارامتر تاوان[19] نامیده می‌شود. اگر λ بی‌نهایت باشد، تنها عرض از مبدأ در مدل حضور خواهد داشت و اگر برابر صفر باشد، قسمت تاوان از مدل حذف‌شده و مدل کمترین توان‌های دوم حاصل خواهد شد.

در اصل، λ و t نقش یکسانی در مدل ایفا می‌کنند و کنترل بزرگی یا کوچکی ناحیه محدودیت (تاوان) بر دوش این دو پارامتر است. جالب‌توجه است که رابطه این دو پارامتر عکس یکدیگر است.

لازم به ذکر است که در دو عبارت (۳) و (۴) از تابع زیان توان دوم خطا[20] استفاده‌شده است که بیانگر روش کمترین توان‌های دوم تاوانیده است. اما به‌طور کلی مسئله بهینه‌سازی تاوانیده به‌صورت

$$\hat{\beta}=\min_{\beta\in\mathbb{R}^p}\Big\{L(\mathbf{y},\mathbf{X}\beta)+\lambda p(\beta)\Big\}$$

نوشته می‌شود که در آن $L(\mathbf{y},\mathbf{X}\beta)$ تابع زیان نامنفی برای نیکویی برازش، $p(\beta)$ تابع تاوان نامنفی و λ پارامتر منظم‌سازی است که تعادل بین نیکویی برازش و پیچیدگی مدل (حضور متغیرهای زیاد) را برقرار می‌کند [۱۹].

در ادامه به معرفی چندین روش تاوانیده پرداخته می‌شود.

۴.۲ رگرسیون بریج

نخستین بار فرانک و فریدمن [۱۴] رگرسیون بریج[21] را معرفی نمودند. آن‌ها در این روش، تابع هدف را با در نظر گرفتن محدودیت $\sum_{j=1}^{p}|\beta_j|^\gamma\leq t$ کمینه و مسئله بهینه‌سازی را به‌صورت

$$\min_{\beta\in\mathbb{R}^p}\Big\{\sum_{i=1}^{n}(y_i-\sum_{j=1}^{p}\mathbf{X}_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}|\beta_j|^\gamma\Big\},\quad i=1,\dots,n,\quad j=1,\dots,p$$

تدوین کردند که در آن γ مقداری نامنفی است. البته فرانک و فریدمن [۱۴] این مسئله بهینه‌سازی را برای تمامی مقادیر ممکن γ به کار نگرفتند. لازم به ذکر است که به ازای $\gamma>1$ ضرایب روش بریج معمولاً برای تمامی متغیرها غیر صفر هستند.

خانواده بریج شامل موارد خاص رگرسیون بهترین زیرمجموعه به ازای $\gamma=0$، رگرسیون لاسو به ازای $\gamma=1$ و رگرسیون ستیغی به ازای $\gamma=2$ می‌باشد که در شکل ۱ رگرسیون بریج به ازای این مقادیر رسم شده است.

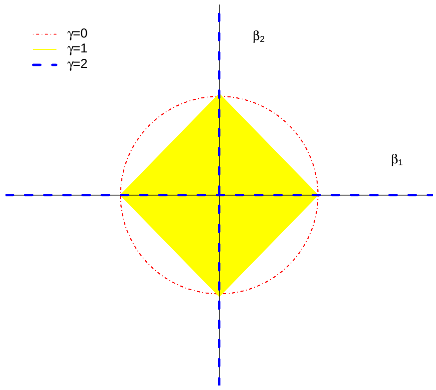
یک ویژگی بسیار سودمندی که برخی از روش‌های تاوانیده دارند، محدب بودن مسئله بهینه‌سازی آن‌هاست. مسئله بهینه‌سازی روش بریج به ازای $\gamma\geq1$ محدب و به ازای $\gamma<1$ غیر محدب است [۱۹].

۵.۲ رگرسیون ستیغی

آندری نیکولایویچ تیخونوف[22] متخصص ژئوفیزیک و ریاضی‌دان روسی که در زمینه کاربردهای ریاضی در فیزیک کار می‌کرد، منظم‌سازی تیخونوف را به‌عنوان راه‌حلی برای رویارویی با مسائل بدشرطیده معرفی نمود. البته او این روش را علاوه بر مسائل بدشرطیده برای موضوعات بسیاری نظیر معادلات انتگرال، توپولوژی، تحلیل عملکردی و فیزیک ریاضی به کار برد. پس از او دیوید فیلیپس[23] به‌طور گسترده‌ای از این روش استفاده کرد. بدین‌سان در برخی از منابع، از این روش، تحت عنوان منظم‌سازی تیخونوف-فیلیپس یادشده است. نخستین بار هورل [۱۶] رگرسیون ستیغی را به‌عنوان یک روش انتخاب متغیر در آمار، به رسمیت شناخت.

برای محاسبه برآوردگر کمترین توان‌های دوم یعنی $\hat{\beta}_{\mathrm{OLS}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ نیاز به محاسبه وارون ماتریس $\mathbf{X}^\top\mathbf{X}$ است. ولی این ماتریس در مجموعه داده‌های با بعد بالا به سبب $p>>n$ رتبه کامل نبوده و درنتیجه وارون‌پذیر نخواهد بود. البته می‌توان برای یافتن وارون ماتریس $\mathbf{X}^\top\mathbf{X}$ از روش وارون تعمیم‌یافته نیز استفاده نمود اما در این روش وارون به‌دست‌آمده منحصربه‌فرد نبوده و به دنبال آن پاسخ‌های به‌دست‌آمده نیز منحصربه‌فرد نخواهد بود. گاهی اوقات وارون وجود دارد ولی کران‌دار نیست و بدین سبب لازم است از روش‌های دیگری استفاده شود. روش ستیغی مقدار مثبتی مانند λ را به درایه‌های قطر اصلی ماتریس $\mathbf{X}^\top\mathbf{X}$ می‌افزاید و با انجام این عمل می‌توان وارون ماتریس را به‌سادگی یافت. البته این امر موجب تحریف داده‌ها

[17] Penalty function
[18] Regularization parameter
[19] Penalty parameter
[20] Square error loss function
[21] Bridge regression
[22] Andrey Nikolayevich Tikhonov
[23] David L. Phillips



شکل ۱: ناحیه محدودیت بریج.

می‌شود که یکی از معایب بزرگ این روش است. بزرگ‌ترین مزیت روش ستیغی وارون‌پذیری همیشگی ماتریس $\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p$ است. سؤال مهمی که در این قسمت پیش می‌آید، این است که چرا در روش ستیغی، وارون همیشه قابل‌محاسبه است؟ برای پاسخ به این سؤال از تجزیه مقادیر منفرد[24] ماتریس‌ها استفاده می‌کنیم. هر ماتریسی حتی ماتریس‌های غیرمربعی دارای تجزیه مقادیر منفرد هستند. این روش یک ماتریس را به سه ماتریس دیگر تجزیه می‌کند. به‌عنوان‌مثال ماتریس $\mathbf{G}$ را می‌توان به‌صورت $\mathbf{G}=\mathbf{UDV}$ تجزیه نمود. به‌طوری‌که $\mathbf{U}$ و $\mathbf{V}$ ماتریس‌های متعامد و $\mathbf{D}$ ماتریس قطری است. البته اگر ماتریس $\mathbf{G}$ متقارن باشد، آنگاه $\mathbf{V}=\mathbf{U}^\top$ به‌صورت قطری متعامد[25] است [۲۶]. به همین ترتیب ماتریس $\mathbf{X}^\top\mathbf{X}$ نیز دارای تجزیه SVD است. از طرفی این ماتریس متقارن است در نتیجه به‌صورت متعامد قطری‌پذیر است. بنابراین می‌توان این ماتریس را به‌صورت

$$\mathbf{X}^\top\mathbf{X}=\mathbf{VDV}^\top$$

نوشت، که در آن $\mathbf{D}$ یک ماتریس قطری است که عناصر روی قطر آن مقادیر ویژه‌ی ماتریس $\mathbf{X}^\top\mathbf{X}$ بوده و $\mathbf{V}$ ماتریس متعامد است که ستون‌های آن بردارهای ویژه $\mathbf{X}^\top\mathbf{X}$ است. از سوی دیگر ماتریس $\mathbf{X}^\top\mathbf{X}$ نیمه معین مثبت است زیرا

$$\forall \mathbf{z}\neq 0,\quad \mathbf{z}^\top\mathbf{X}^\top\mathbf{X}\mathbf{z}=(\mathbf{X}\mathbf{z})^\top(\mathbf{X}\mathbf{z})=||\mathbf{X}\mathbf{z}||^2\geq 0$$

به همین جهت عناصر قطری بزرگ‌تر یا مساوی صفر هستند. اکنون با نمایش $\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p$ به‌صورت $\mathbf{VDV}^\top+\mathbf{V}\lambda\mathbf{I}_p\mathbf{V}^\top$ واضح است که $\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p$ قطری متعامد است. بنابراین

$$\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}_p=\mathbf{V}(\mathbf{D}+\lambda\mathbf{I}_p)\mathbf{V}^\top$$

به‌طوری‌که تمام مقادیر ویژه آن (با توجه به این‌که $\lambda>0$ است) بزرگ‌تر از صفرند و زمانی که تمام مقادیر ویژه مخالف صفر باشند، ماتریس مربوطه وارون‌پذیر است. مسئله بهینه‌سازی روش ستیغی به‌صورت زیر است:

$$\begin{aligned}&\min_{\beta\in\mathbb{R}^p}\left\{(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta)+\lambda\beta^\top\beta\right\}\\&=\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}(y_i-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right\}\end{aligned}$$

که در آن تابع تاوان روش ستیغی $\sum_{j=1}^{p}\beta_j^2$ می‌باشد. به λ، پارامتر ستیغی[26] نیز گفته می‌شود و برخی آن را با k نیز نمایش می‌دهند. مسئله بهینه‌سازی روش ستیغی محدب است و بدین‌سان موردعلاقه محققین است.

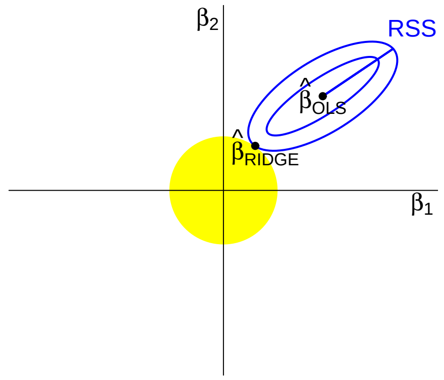
برای یافتن درک بصری از روش ستیغی با فرض وجود تنها دو متغیر توضیحی در مدل رگرسیونی ناحیه محدودیت به‌صورت

$$\beta_1^2+\beta_2^2\leq t \qquad (۵)$$

است. عبارت (۵) نشان‌دهنده معادله دایره‌ای به مرکز مبدأ مختصات و شعاع $\sqrt{t}$ می‌باشد. بنابراین ناحیه‌ی محدودیت یک دایره است.

---

[24]Singular value decomposition(SVD)
[25]Orthogonal diagonalization
[26]Ridge parameter



شکل ۲: ناحیه محدودیت ستیغی.

در شکل ۲ دایره زرد رنگ ناحیه تاوان است. محل برخورد RSS (مجموع توان‌های دوم باقی‌مانده‌ها) روش کمترین توان‌های دوم با دایره، مختصات برآوردگر ستیغی را نمایش می‌دهد. با عمود کردن این نقطه بر محور افقی، $\hat{\beta}_1$ و با عمود کردن آن بر محور عمودی، $\hat{\beta}_2$ حاصل می‌شود. روش ستیغی فرآیندی پیوسته است که ضرایب تولیدی آن پایدار هستند اما برآوردهایی اریب ارائه می‌دهد که پذیرفتن مقداری اریبی در مقابل یافتن پاسخی قابل‌اطمینان‌تر معقول به نظر می‌رسد.

روش ستیغی ضرایب را به سمت صفر منقبض می‌کند، هرچه مقدار پارامتر ستیغی بیشتر باشد ضرایب برآورد شده انقباض بیشتری داشته و کوچک‌تر می‌شوند ولی هیچ‌گاه حتی در بی‌نهایت نیز به صفر نمی‌رسند این امر به سبب خاصیت دایره (ناحیه تاوان روش ستیغی) رخ می‌دهد. البته به ازای $p > ۲$ نیز هیچ‌گاه ضرایب برآورد شده به صفر نمی‌رسند. بدین‌سان به‌کارگیری این روش برای داده‌های بعد بالا، مدل‌های قابل تفسیر ارائه نمی‌دهد و این روش تنک نیست.

۶.۲ رگرسیون لاسو

رگرسیون لاسو[27] توسط رابرت تیبشیرانی[28] در سال ۱۹۹۶ پیشنهاد شد. ایده اولیه لاسو برگرفته از پیشنهاد جالب لئو برایمن[29] بود [۲۳]. مسئله بهینه‌سازی نامنفی برایمن که گروتی[30] نام دارد، به‌صورت

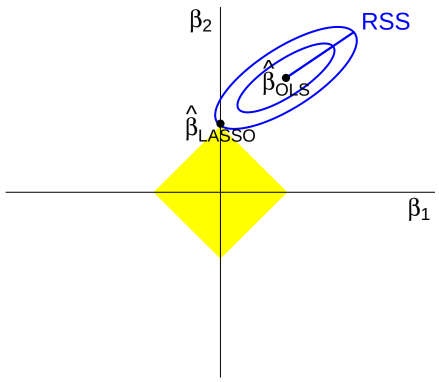
$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} c_j\hat{\beta}_j x_{ji}\right)^2 \qquad c_j \geq 0, \quad \sum_j c_j \leq s$$

است که در آن $\hat{\beta}_j$ برآوردگر اولیه کمترین توان‌های دوم است. با کاهش s (پارامتر منظم‌سازی) گروتی محدود می‌شود. روش گروتی برخی از متغیرها را حذف می‌کند و مابقی آن‌ها را منقبض می‌کند. بنابراین این روش مدل‌های قابل تفسیر ارائه می‌دهد. این روش نسبتاً پایدار است و همچنین مقیاس پایاست [۲۳]. مطالعات برایمن [۹] روی داده‌های واقعی و داده‌های شبیه‌سازی‌شده نشان داد که خطای پیش‌گویی گروتی در مقایسه با رگرسیون بهترین زیرمجموعه کمتر است و در رقابتی پایاپای با رگرسیون ستیغی است، به‌جز زمانی که مدل واقعی ضرایب بسیار کوچک غیر صفر دارد. وابستگی گروتی به علامت و مقدار برآورد کمترین توان‌های دوم نقص بزرگی است که بی‌اعتمادی محققان به این روش را به ارمغان آورده است. بدین‌جهت می‌توان گفت زمانی که با داده‌های بعد بالا مواجه هستیم و یا زمانی که متغیرها بسیار همبسته‌اند، برآوردهای کمترین توان‌های دوم عملکرد ضعیفی دارد و به‌تبع آن گروتی نیز عملکرد ضعیفی داشته و مورداطمینان نیست. این در حالی است که لاسو برخلاف گروتی از استفاده مستقیم از برآوردگرهای کمترین توان‌های دوم اجتناب می‌کند. مسئله بهینه‌سازی لاسو به‌صورت

$$\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\},$$

[27] Lasso
[28] Robert Tibishirani
[29] Leo Breiman
[30] Garrote

است. تیبشیرانی [۲۳] برای حل مسئله بهینه‌سازی بالا یک مسئله برنامه‌ریزی درجه دوم با محدودیت نابرابری خطی به‌کاربرده است. این محدودیت به‌طور طبیعی تمایل به تولید ضرایبی که صفر هستند، دارد و ازاین‌رو مدل‌های قابل تفسیر ارائه می‌دهد. روش لاسو برآوردگرهای پایدار، اریب و پیوسته تولید می‌کند [۲۳]. یعنی برآوردگر تولیدشده روی داده‌ها پیوسته است و ناپایداری در مدل را کاهش می‌دهد. ناپایداری در مدل، یعنی با ایجاد کوچک‌ترین تغییراتی نتیجه انتخاب مدل بسیار متفاوت خواهد شد و دقت نتایج کاهش میابد. لاسو به معنی عملگر انتخاب و کمترین قدرمطلق انقباضی[۳۱] است. لاسو برخی ضرایب را منقبض و بقیه را صفر می‌کند. بدین سبب هر دو ویژگی خوب رگرسیون انتخاب زیرمجموعه و ستیغی را حفظ می‌کند.



شکل ۳: ناحیه محدودیت لاسو.

برای یافتن درک بصری از روش لاسو با فرض وجود تنها دو متغیر توضیحی در مدل رگرسیونی، ناحیه محدودیت به‌صورت

$$|\beta_۱|+|\beta_۲|\leq t$$

است که با رسم آن، ناحیه محدودیت یک لوزی است. در شکل ۳ لوزی زرد رنگ ناحیه تاوان است. محل برخورد RSS روش کمترین توان‌های دوم با لوزی، مختصات برآوردگر لاسو را نمایش می‌دهد. با عمود کردن این نقطه بر محور افقی، $\hat{\beta}_۱$ و با عمود کردن آن بر محور عمودی، $\hat{\beta}_۲$ حاصل می‌شود. اگر محل برخورد دقیقاً روی یکی از گوشه‌های لوزی باشد، (همانند شکل ۳) تنها یک متغیر انتخاب‌شده و دیگری صفر می‌شود. همان‌طور که ذکر شد، در روش لاسو به سبب نوع ناحیه تاوان (لوزی و گوشه داشتن لوزی) برخی از متغیرها دقیقاً برابر صفر می‌شوند. از منظر دیگر می‌توان استدلال کرد که چون تابع تاوان روش لاسو در نقطه صفر مشتق ناپذیر است، مقادیر بزرگ‌تری از λ منجر به حذف متغیرهای پیشگوی کم تأثیر در مدل می‌گردد و در نتیجه عمل برآورد پارامتر و انتخاب متغیر به‌صورت هم‌زمان انجام می‌گیرد [۳].

۷.۲ روش شبکه ارتجاعی

این روش که ترکیبی از روش لاسو و ستیغی است توسط ژو و هستی [۲۷] معرفی شد. آن‌ها این روش را برای تعدیل هم‌زمان مشکلات هم‌خطی چندگانه و تفسیر ناپذیر بودن مدل، مطرح کردند.

مسئله بهینه‌سازی به‌صورت

$$\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=۱}^{n}(y_i-\sum_{j=۱}^{p}\mathbf{X}_{ij}\beta_j)^۲+\lambda_۱\sum_{j=۱}^{p}|\beta_j|+\lambda_۲\sum_{j=۱}^{p}\beta_j^۲\right\}$$

روش شبکه ارتجاعی[۳۲] دو پارامتر تاوان $\lambda_۱$ و $\lambda_۲$ را داراست که طبیعتاً یافتن مقدار مناسب برای دو پارامتر تاوان این روش نسبت به روش‌هایی که تنها یک پارامتر تاوان دارند، امری به‌مراتب دشوارتر و سخت‌تر خواهد بود و این‌یکی از معایب این روش محسوب می‌شود. تابع تاوان این روش، محدب است.

[31] Least Absolute Shrinkage and Selection Operator (LASSO)
[32] Elastic Net

۸.۲ روش اسکد

فن و لی [۱۲] استفاده از روش اسکد[33] را که تابع تاوان آن توسط [۱۳] ارائه‌شده بود، پیشنهاد کردند. تابع تاوان پیشنهادی آن‌ها به‌صورت زیر است:

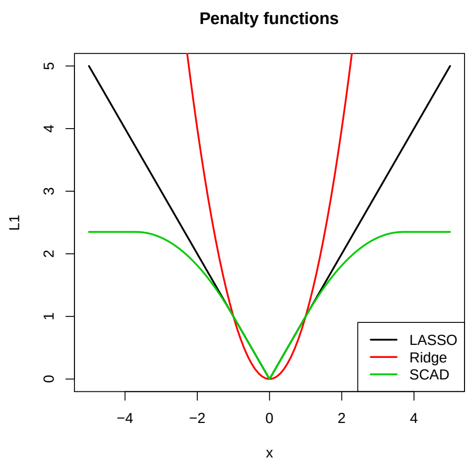
$$P_\lambda(\beta_j)=\begin{cases}\lambda|\beta_j| & \text{اگر} \quad |\beta_j|\le\lambda;\\ -\left(\frac{|\beta_j|^2-2a\lambda|\beta_j|+\lambda^2}{2(a-1)}\right) & \text{اگر} \quad \lambda<|\beta_j|\le a\lambda;\\ \frac{(a+1)\lambda^2}{2} & \text{اگر} \quad |\beta_j|>a\lambda\end{cases}$$

λ و a دو پارامتر تاوان این روش هستند که همین امر سبب دشواری این روش است. (پارامتر a را با γ نیز نمایش می‌دهند.) برآوردهای این روش ناریب و پیوسته هستند. در این روش نیز متغیرهای کم تأثیر از مدل خارج می‌شوند. به بیان ساده‌تر این روش تنک است. یکی از عیوب بزرگ این روش غیرمحدب بودن مسئله بهینه‌سازی آن است. فن و لی [۱۲] مقدار $a=3/7$ را به‌عنوان مقداری مناسب برای حل مسئله بهینه‌سازی این روش پیشنهاد نمودند. در شکل ۴ تابع تاوان روش لاسو، ستیغی و اسکد به‌طور هم‌زمان رسم شده‌اند.

۹.۲ روش اسکار

باندل و ریچ [۸] روش اسکار[34] را منتشر کردند. این روش در صورت وجود هم‌خطی عملکرد خوبی دارد. ناحیه تاوان این روش به‌صورت زیر است.

$$\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^n\left(y_i-\sum_{j=1}^p \mathbf{X}_{ij}\beta_j\right)^2 + \lambda\left(\sum_{j=1}^p|\beta_j|+c\sum_{j<k}\max\{|\beta_j|,|\beta_k|\}\right)\right\} \quad c\ge 0.$$



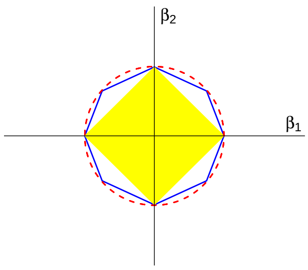
شکل ۴: تابع تاوان سه روش لاسو، ستیغی و اسکد.

که در آن پارامتر c یک ثابت تنظیم‌کننده است. تاوان $\sum_{j=1}^p|\beta_j|$ باعث تنکی و همچنین تاوان $\sum_{j<k}\max\{|\beta_j|,|\beta_k|\}$ باعث تساوی ضرایب می‌شود. تابع هدف این روش نیز محدب است. هشت‌ضلعی موجود در شکل ۵ ناحیه تاوان این روش است.

خانواده‌ی برآوردگرهای کمترین توان‌های دوم تاوانیده[35] خانواده‌ی وسیعی از برآوردگرها است که توصیف و ذکر تمامی این موارد در این مقاله نمی‌گنجد.

یکی از معایب بزرگ روش‌های تاوانیده، وابستگی آن‌ها به پارامتر تاوان است. افزایش پارامتر تاوان، افزایش اریبی و کاهش واریانس را به همراه دارد. همچنین باعث بزرگی ناحیه محدودیت شده که این امر در روش‌های تنک منجر به حذف متغیرهای فراوانی می‌شود. بدین‌سان حتی ممکن است متغیرهای پراهمیت از مدل حذف شود به‌طوری‌که در بی‌نهایت تمامی متغیرها را از مدل خارج می‌کند. کاهش این پارامتر، نیز کاهش اریبی و افزایش واریانس را به ارمغان می‌آورد. همچنین باعث کوچکی ناحیه محدودیت می‌شود. بدین‌جهت ممکن است متغیرهای کم‌اهمیت زیادی در مدل باقی

[33] Smoothly Clipped Absolute Devaition (SCAD)
[34] Octagonal shrinkage and clustering algorithm for regression (OSCAR)
[35] The penalized least squares(PLS) family



شکل ۵: ناحیه محدودیت اسکار.

بمانند. با انتخاب صفر برای این پارامتر روش تاوانیده دقیقاً معادل روش کمترین توان‌های دوم خواهد بود. بدین منظور یکی از نکات بسیار مهم در روش تاوانیده تعیین مقدار مناسبی برای پارامتر تاوان است. منظور از مقدار مناسب، مقداری است که توازنی بین اریبی و واریانس برقرار نماید، به‌طوری‌که RSS مدل را به کمترین مقدار ممکن برساند. برای تعیین این مقدار بهینه روش‌هایی نظیر معیار اطلاع آکائیکه[۳۶]، اطلاع بیز[۳۷]، آکائیکه تصحیح‌شده، C_p مالوس، اعتبارسنجی[۳۸] و ... وجود دارد. اخیراً رایج‌ترین روش در میان محققان برای محاسبه‌ی این پارامتر، روش اعتبارسنجی است. دلیل محبوبیت این روش، محاسبه‌ی هم‌زمان مقدار بهینه و آزمون میزان دقت آن می‌باشد.

۱۰.۲ تابع تاوان

در بخش قبل برخی از روش‌های تاوانیده به‌اختصار بیان شد. اما واقعیت این است که توابع تاوان فراوانی وجود دارد که هر کدام منجر به ایجاد روش جدیدی می‌شوند. سؤال مهمی که در این قسمت پیش می‌آید، این است که کدام از این توابع تاوان موردپسند هستند و یا این‌که ویژگی‌های مطلوب یک تابع تاوان چیست؟ فن و لی [۱۲] با انتشار مقاله‌ای به این سؤال‌ها پاسخ دادند. از نظر آن‌ها یک تابع تاوان مطلوب باید هم‌زمان سه ویژگی مهم که به شرح ذیل است، را دارا باشد.

۱. تنکی: تابع تاوان موردنظر باید ضرایبی که مقدارشان کوچک است را برابر صفر قرار دهد. همین امر منجر به کاهش پیچیدگی و افزایش تفسیرپذیری مدل خواهد شد.

۲. نااریبی: تابع تاوان موردنظر باید منجر به ایجاد برآوردی نااریب شود، یعنی

$$E(\hat{\beta}) = \beta$$

۳. پیوستگی: تابع توان موردنظر منجر به ایجاد روشی پایدار و برآوردگرهایی پیوسته شود.

یکی دیگر از موارد بسیار مهم در روش تاوانیده که رویه انتخاب متغیر دارند، این است که بتوانند زیرمجموعه درست و واقعی متغیرها را شناسایی کنند. این خاصیت الهام‌بخش، تحت عنوان خاصیت پیشگویی[۳۹] شناخته می‌شود که برای نخستین بار توسط فن و لی [۱۲] مطرح شد.

لازم به ذکر است که بررسی این خاصیت در عمل برای داده‌های شبیه‌سازی‌شده امکان‌پذیر است.

در میان تمامی روش‌های مطرح‌شده، روش‌هایی که تابع تاوان آن‌ها محدب است، موردپسندتر است. چراکه اثبات می‌شود در مسائل بهینه‌سازی محدب[۴۰]، مینیمم (ماکزیمم) نسبی با مینیمم (ماکزیمم) مطلق معادل است.

۱۱.۲ روش تاوانیده اصلاح‌شده

همان‌طور که مطرح شد روش‌های تاوانیده دارای معایبی نظیر انتخاب مقدار بهینه‌ی پارامتر تاوان و ... است که نقش مهمی در نتیجه‌ی برآورد دارد، به این معنی که مقادیر بزرگ پارامتر تاوان منجر به بی‌ثباتی قابل‌توجهی می‌شود. انتخاب مقدار پارامتر تاوان باوجود تمامی تحقیقات انجام‌شده و روش‌های متنوع پیشنهادشده توسط محققین رشته‌های آمار و ریاضی هنوز به‌طور کامل حل‌نشده است و همچنان عرصه برای تلاش بیشتر و معرفی روش‌های دقیق‌تر

[36] Akaike information criterion (AIC)
[37] Bayesian information criterion (BIC)
[38] Cross-validation(CV)
[39] Oracle property
[40] Convex optimazation

به‌منظور انتخاب بهینه‌ترین مقدار وجود دارد. زیرا برآورد به‌دست‌آمده در روش‌های تاوانیده وابستگی زیادی به پارامتر تاوان مجهول دارد.

در واقع مشکل بنیادینی که داده‌ی بعد بالا برای ما به ارمغان می‌آورد عدم وارون‌پذیری ماتریس $\mathbf{X}^\top\mathbf{X}$ به سبب بدشرطیده[41] بودن آن است. ماتریس بدشرطیده دارای عدد شرطی[42] بزرگی است. به همین شکل برای مقابله با این مشکل تنها کافی است ماتریس $\mathbf{X}^\top\mathbf{X}$ با کاهش عدد شرطی به یک ماتریس خوش‌شرطیده[43] تبدیل شود. به همین سبب کافی است که ضریبی مثبت از عدد شرطی ماتریس $\mathbf{X}^\top\mathbf{X}$ به تابع هدف به‌عنوان عبارت تاوان، به‌منظور کنترل عدد حالت ماتریس افزوده می‌شود. بنابراین با استفاده از موارد ذکرشده و استفاده از روش پیشنهادی روزبه و همکاران [۲۲] می‌توان نوع اصلاح‌شده مسئله بهینه‌سازی (۴) را به‌صورت

$$\min_{\beta\in\mathbb{R}^p}\left\{(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta)+M\mathscr{K}(\mathbf{X}^\top\mathbf{X})\right\},$$

نوشت، که در آن $M\geq ۰$ پارامتر تاوان نامیده شده و $\mathscr{K}(\mathbf{X}^\top\mathbf{X})$ عدد شرطی ماتریس $\mathbf{X}^\top\mathbf{X}$ است. اعمال این روش موجب رهایی از دردسرهای روش‌های تاوانیده می‌شود، اما متأسفانه تنک نیست. این روش برای مقابله با هم‌خطی نیز سودمند خواهد بود.

۳ مطالعه شبیه‌سازی

در این بخش در یک مطالعه شبیه‌سازی به بررسی و کاربرد روش‌های بیان‌شده می‌پردازیم. در این مطالعه نمونه‌ای با اندازه ۵۰ به همراه ۱۰۰ متغیر توضیحی از مدل

$$\mathbf{y}=\mathbf{X}\beta+\varepsilon$$

تولیدشده‌اند. برای تولید بردار پارامترها از

$$\beta_۱=(-۱/۵,۲,۲/۵,۴,-۳,۵)^\top,\quad \beta_۲\sim N(۰,۰/۰۱),$$

برای تولید ماتریس طرح از

$$\mathbf{X}\sim N_p(\mu,\mathbf{I}),\quad \mu=(۰,\ldots,۰)^\top_{p\times ۱}$$

و برای تولید خطا از

$$\varepsilon\sim N(۰,۱)$$

استفاده شد. در برخی از روش‌های به‌کاررفته در این مثال نیازمند استفاده از روش اعتبارسنجی متقابل هستیم. در تمام این روش‌ها مجموعه آموزش ۷۰٪ و مجموعه آزمون ۳۰٪ در نظر گرفته‌شده است. در گام اول از روش مؤلفه‌های اصلی استفاده می‌شود که نتایج آن در جدول ۱ خلاصه‌شده است. با توجه به جدول ۱ به ازای ۱۷ مؤلفه ۶۲٪ تغییرات، به ازای ۲۱ مؤلفه ۷۱٪ تغییرات، به ازای ۲۶ مؤلفه ۸۰٪ تغییرات، به ازای ۳۴ مؤلفه ۹۰٪ تغییرات، قابل توجیه بوده و از مؤلفه چهل‌وهشتم به بعد ۱۰۰ درصد تغییرات توسط واریانس کل توجیه می‌شود. نمودار بازو و لگاریتم آن در شکل ۶ نمایش داده‌شده است. تعداد مؤلفه‌ها با این نمودار و نمودار لگاریتم ۴۹ مؤلفه تعیین می‌شود. در ادامه از روش‌های تاوانیده استفاده می‌کنیم در شکل ۷ و ۸ نمودارهای مربوط به اعتبارسنجی و برآورد ضرایب به ازای پارامترهای تاوانیده مختلف با روش ستیغی و لاسو مشهود است. در نمودار اعتبارسنجی روش لاسو به ازای مقدار ۰/۱۰۹۷ مقدار MSE را کمینه کرده و در روش لاسو مقدار آن به ازای ۲۳/۱۰۱۳ کمینه شده است. در قسمت فوقانی نمودارهای سمت راست شکل‌های ۷ و ۸ تعداد ضرایب غیر صفر موجود به ازای مقادیر مختلف λ نمایش داده‌شده است که واضح است که با افزایش میزان λ تعداد ضرایب غیر صفر روش لاسو کاهش می‌یابد درحالی‌که در روش ستیغی برآورد ضرایب بسیار به صفر نزدیک می‌شوند ولی دقیقاً معادل با صفر نخواهند بود. نمودار اعتبارسنجی برای روش شبکه ارتجاعی در شکل ۹ نمایش داده‌شده است. نمودارهای برآورد ضرایب به روش شبکه ارتجاعی به ازای مقادیر مختلف پارامتر α و بردار پارامتر دلخواه یکسان در شکل ۱۲ رسم شده است. با توجه به این شکل‌ها بدیهی است که با افزایش میزان α متغیرهای بیشتری از مدل خارج می‌شوند. نمودارهای برآورد ضرایب به روش SCAD به ازای مقادیر مختلف پارامتر a و بردار پارامتر دلخواه یکسان در شکل ۱۰ رسم شده است. نمودار اعتبارسنجی متقابل برای یافت مقدار بهینه پارامتر λ به ازای مقدار پیشنهادی فن و لی [۱۲] برای پارامتر a در شکل ۱۱ نشان داده‌شده است. در جدول ۲ نتایج روش‌های مورداستفاده گزارش‌شده است.

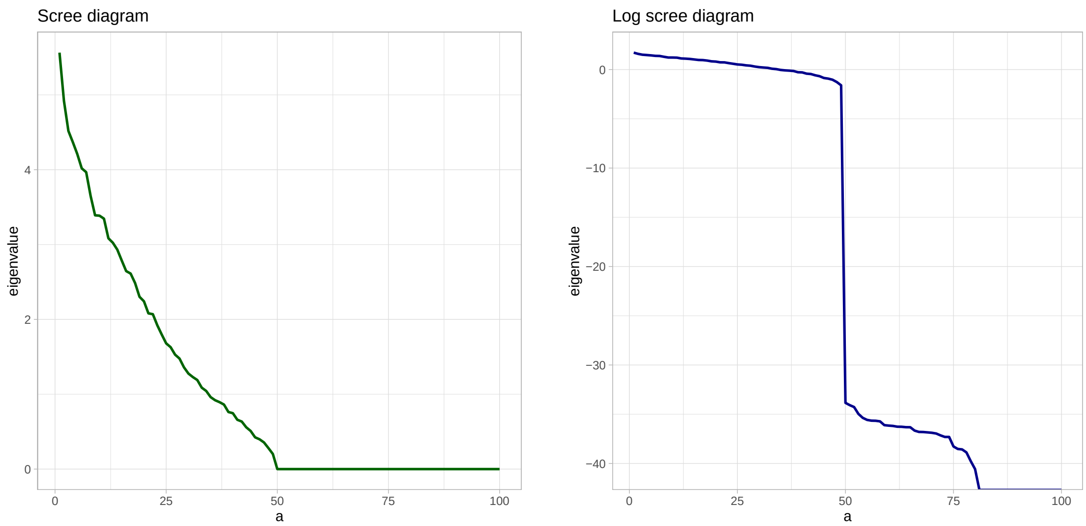
۴ مطالعه‌ی داده واقعی

در این بخش به‌منظور بررسی یک مجموعه داده واقعی با بعد بالا از داده‌های ریبوفلاوین استفاده می‌شود. ریبوفلاوین که به‌عنوان ویتامین B2 نیز شناخته می‌شود، یکی از ویتامین‌های B است که همه محلول در آب هستند. از مهم‌ترین ویتامین‌های لازم برای حیات موجود زنده است. دلیل انتخاب این نام، رنگ زرد این ویتامین است که ناشی از وجود حلقه فلاوینی موجود در ساختمان آن می‌باشد. ریبوفلاوین به‌طور طبیعی در برخی غذاها وجود دارد، به

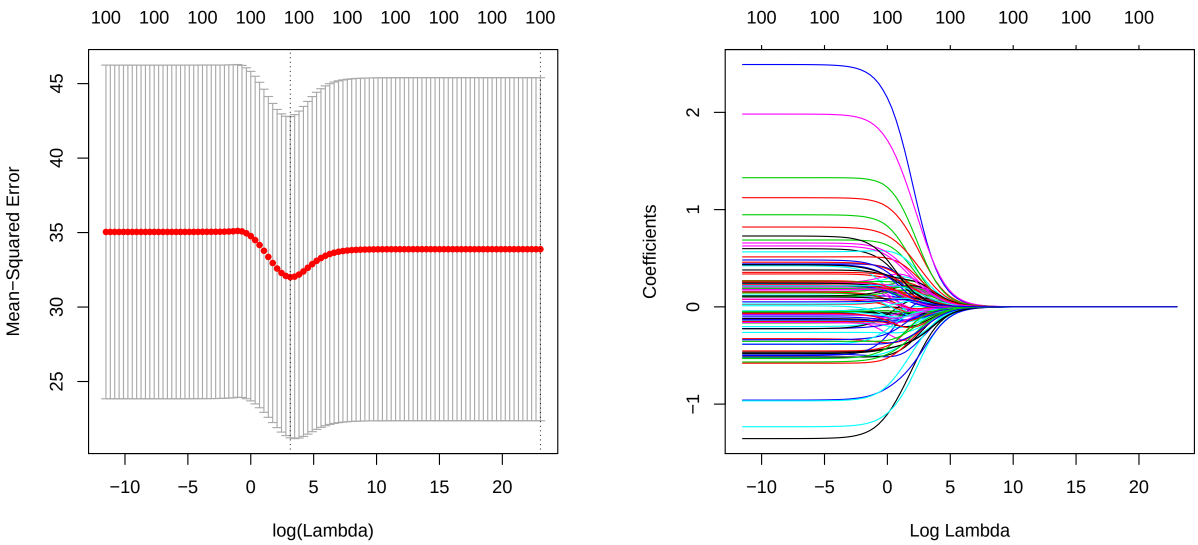
[41] Ill-condition

[42] Condition number

[43] Well-condition



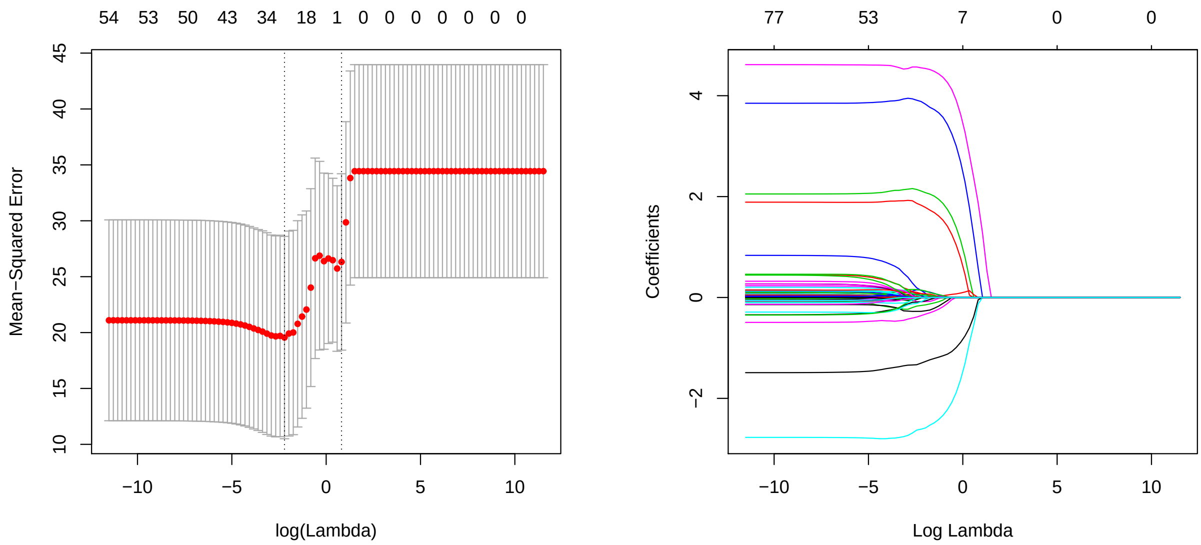
شکل ۶: نمودار بازو و لگاریتم آن برای داده‌های شبیه‌سازی‌شده.



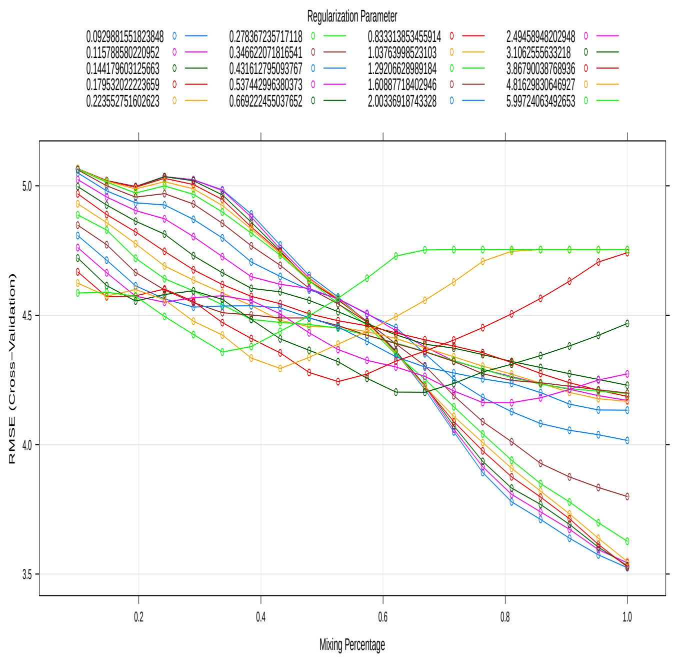
شکل ۷: نمودار اعتبارسنجی و برآورد ضرایب داده‌های شبیه‌سازی‌شده به ازای پارامترهای تاوان مختلف به روش ستیغی.

جدول ۱: نتایج روش مؤلفه‌های اصلی برای داده‌های شبیه‌سازی‌شده.

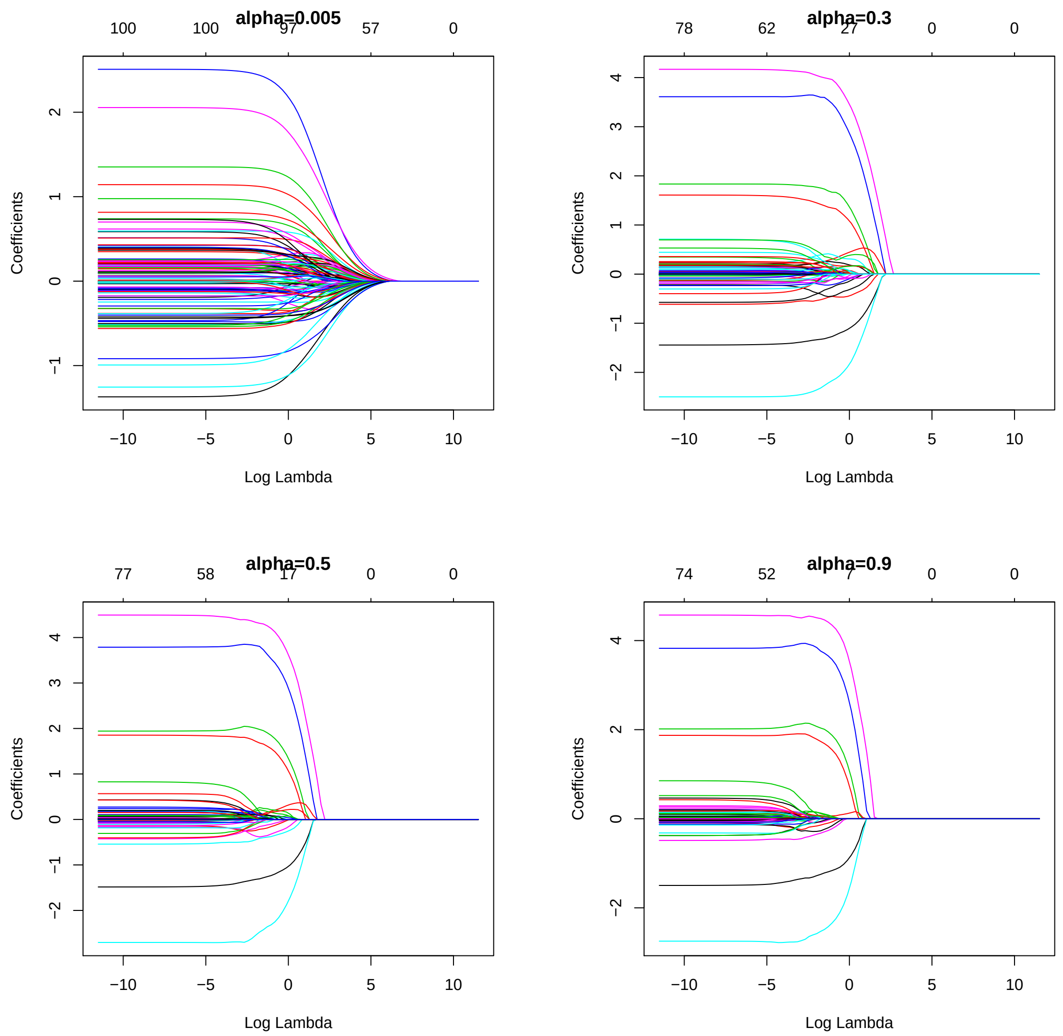
مؤلفه	نسبت واریانس	نسبت تجمعی واریانس	مؤلفه	نسبت واریانس	نسبت تجمعی واریانس
مؤلفه اول	۰/۰۵۵۶	۰/۰۵۵۶	مؤلفه پنجاه‌ویکم	$۱/۵۸۶۹ \times ۱۰^{-۱۷}$	۱
مؤلفه دوم	۰/۰۴۹۱	۰/۱۰۴۸	مؤلفه پنجاه‌ودوم	$۱/۳۰۹۹ \times ۱۰^{-۱۷}$	۱
مؤلفه سوم	۰/۰۴۵۱	۰/۱۵۰۰	مؤلفه پنجاه‌وسوم	$۶/۵۳۳۰ \times ۱۰^{-۱۸}$	۱
مؤلفه چهارم	۰/۰۴۳۶	۰/۱۹۳۶	مؤلفه پنجاه‌وچهارم	$۴/۳۸۲۱ \times ۱۰^{-۱۸}$	۱
مؤلفه پنجم	۰/۰۴۲۰	۰/۲۳۵۷	مؤلفه پنجاه‌وپنجم	$۳/۵۵۷۹ \times ۱۰^{-۱۸}$	۱
مؤلفه ششم	۰/۰۴۰۱	۰/۲۷۵۹	مؤلفه پنجاه‌وششم	$۳/۲۹۰۴ \times ۱۰^{-۱۸}$	۱
مؤلفه هفتم	۰/۰۳۹۶	۰/۳۱۵۶	مؤلفه پنجاه‌وهفتم	$۳/۲۴۹۰ \times ۱۰^{-۱۸}$	۱
مؤلفه هشتم	۰/۰۳۶۴	۰/۳۵۲۱	مؤلفه پنجاه‌وهشتم	$۳/۰۶۱۹ \times ۱۰^{-۱۸}$	۱
مؤلفه نهم	۰/۰۳۳۹	۰/۳۸۶۰	مؤلفه پنجاه‌ونهم	$۲/۰۸۶۴ \times ۱۰^{-۱۸}$	۱
مؤلفه دهم	۰/۰۳۳۸	۰/۴۱۹۹	مؤلفه شصتم	$۱/۹۹۲۰ \times ۱۰^{-۱۸}$	۱
مؤلفه یازدهم	۰/۰۳۳۴	۰/۴۵۳۳	مؤلفه شصتم‌ویکم	$۱/۹۲۳۶ \times ۱۰^{-۱۸}$	۱
مؤلفه دوازدهم	۰/۰۳۰۸	۰/۴۸۴۲	مؤلفه شصتم‌ودوم	$۱/۷۸۱۲ \times ۱۰^{-۱۸}$	۱
مؤلفه سیزدهم	۰/۰۳۰۲	۰/۵۱۴۴	مؤلفه شصتم‌وسوم	$۱/۷۶۸۶ \times ۱۰^{-۱۸}$	۱
مؤلفه چهاردهم	۰/۰۲۹۳	۰/۵۴۳۷	مؤلفه شصتم‌وچهارم	$۱/۶۹۴۰ \times ۱۰^{-۱۸}$	۱
مؤلفه پانزدهم	۰/۰۲۷۸	۰/۵۷۱۶	مؤلفه شصتم‌وپنجم	$۱/۶۸۵۶ \times ۱۰^{-۱۸}$	۱
مؤلفه شانزدهم	۰/۰۲۶۴	۰/۵۹۸۱	مؤلفه شصتم‌وششم	$۱/۱۹۷۹ \times ۱۰^{-۱۸}$	۱
مؤلفه هفدهم	۰/۰۲۶۱	۰/۶۲۴۲	مؤلفه شصتم‌وهفتم	$۱/۰۴۹۱ \times ۱۰^{-۱۸}$	۱
مؤلفه هجدهم	۰/۰۲۴۸	۰/۶۴۹۱	مؤلفه شصتم‌وهشتم	$۱/۰۳۸۱ \times ۱۰^{-۱۸}$	۱
مؤلفه نوزدهم	۰/۰۲۳۰	۰/۶۷۲۱	مؤلفه شصتم‌ونهم	$۹/۹۸۹۳ \times ۱۰^{-۱۹}$	۱
مؤلفه بیستم	۰/۰۲۲۴	۰/۶۹۴۵	مؤلفه هفتادم	$۹/۶۰۳۸ \times ۱۰^{-۱۹}$	۱
مؤلفه بیست‌ویکم	۰/۰۲۰۸	۰/۷۱۵۳	مؤلفه هفتادویکم	$۸/۹۱۵۷ \times ۱۰^{-۱۹}$	۱
مؤلفه بیست‌ودوم	۰/۰۲۰۶	۰/۷۳۶۰	مؤلفه هفتادودوم	$۷/۳۷۷۴ \times ۱۰^{-۱۹}$	۱
مؤلفه بیست‌وسوم	۰/۰۱۹۲	۰/۷۵۵۲	مؤلفه هفتادوسوم	$۶/۲۹۱۲ \times ۱۰^{-۱۹}$	۱
مؤلفه بیست‌وچهارم	۰/۰۱۷۹	۰/۷۷۳۲	مؤلفه هفتادوچهارم	$۶/۲۹۱۱ \times ۱۰^{-۱۹}$	۱
مؤلفه بیست‌وپنجم	۰/۰۱۶۷	۰/۷۹۰۰	مؤلفه هفتادوپنجم	$۲/۳۹۰۴ \times ۱۰^{-۱۹}$	۱
مؤلفه بیست‌وششم	۰/۰۱۶۲	۰/۸۰۶۳	مؤلفه هفتادوششم	$۱/۸۵۸۷ \times ۱۰^{-۱۹}$	۱
مؤلفه بیست‌وهفتم	۰/۰۱۵۲	۰/۸۲۱۶	مؤلفه هفتادوهفتم	$۱/۷۷۴۰ \times ۱۰^{-۱۹}$	۱
مؤلفه بیست‌وهشتم	۰/۰۱۴۷	۰/۸۳۶۳	مؤلفه هفتادوهشتم	$۱/۳۰۸۹ \times ۱۰^{-۱۹}$	۱
مؤلفه بیست‌ونهم	۰/۰۱۳۵	۰/۸۴۹۹	مؤلفه هفتادونهم	$۵/۳۸۹۹ \times ۱۰^{-۲۰}$	۱
مؤلفه سی‌ام	۰/۰۱۲۷	۰/۸۶۲۷	مؤلفه هشتادم	$۲/۴۰۳۴ \times ۱۰^{-۲۰}$	۱
مؤلفه سی‌ویکم	۰/۰۱۲۲	۰/۸۷۵۰	مؤلفه هشتادویکم	۰	۱
مؤلفه سی‌ودوم	۰/۰۱۱۹	۰/۸۸۶۹	مؤلفه هشتادودوم	۰	۱
مؤلفه سی‌وسوم	۰/۰۱۰۸	۰/۸۹۷۸	مؤلفه هشتادوسوم	۰	۱
مؤلفه سی‌وچهارم	۰/۰۱۰۴	۰/۹۰۸۲	مؤلفه هشتادوچهارم	۰	۱
مؤلفه سی‌وپنجم	۰/۰۰۹۶	۰/۹۱۷۸	مؤلفه هشتادوپنجم	۰	۱
مؤلفه سی‌وششم	۰/۰۰۹۲	۰/۹۲۷۰	مؤلفه هشتادوششم	۰	۱
مؤلفه سی‌وهفتم	۰/۰۰۸۹	۰/۹۳۶۰	مؤلفه هشتادوهفت	۰	۱
مؤلفه سی‌وهشتم	۰/۰۰۸۶	۰/۹۴۴۶	مؤلفه هشتادوهشتم	۰	۱
مؤلفه سی‌ونهم	۰/۰۰۷۶	۰/۹۵۲۲	مؤلفه هشتادونهم	۰	۱
مؤلفه چهلم	۰/۰۰۷۴	۰/۹۵۹۷	مؤلفه نود	۰	۱
مؤلفه چهل‌ویکم	۰/۰۰۶۵	۰/۹۶۶۳	مؤلفه نودویکم	۰	۱
مؤلفه چهل‌ودوم	۰/۰۰۶۳	۰/۹۷۲۶	مؤلفه نودودوم	۰	۱
مؤلفه چهل‌وسوم	۰/۰۰۵۵	۰/۹۷۸۲	مؤلفه نودوسوم	۰	۱
مؤلفه چهل‌وچهارم	۰/۰۰۵۰	۰/۹۸۳۳	مؤلفه نودوچهارم	۰	۱
مؤلفه چهل‌وپنجم	۰/۰۰۴۲	۰/۹۸۷۶	مؤلفه نودوپنجم	۰	۱
مؤلفه چهل‌وششم	۰/۰۰۳۹	۰/۹۹۱۶	مؤلفه نودوششم	۰	۱
مؤلفه چهل‌وهفتم	۰/۰۰۳۵	۰/۹۹۵۱	مؤلفه نودوهفتم	۰	۱
مؤلفه چهل‌وهشتم	۰/۰۰۲۸	۰/۹۹۷۹	مؤلفه نودوهشتم	۰	۱
مؤلفه چهل‌ونهم	۰/۰۰۲۰	۱	مؤلفه نودونهم	۰	۱
مؤلفه پنجاهم	$۲/۰۰۵۵ \times ۱۰^{-۱۷}$	۱	مؤلفه صدم	۰	۱



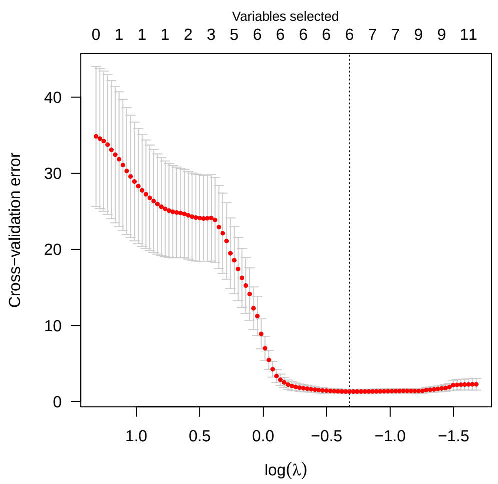
شکل ۸: نمودار اعتبارسنجی و برآورد ضرایب داده‌های شبیه‌سازی‌شده به ازای پارامترهای تاوان مختلف به روش لاسو.



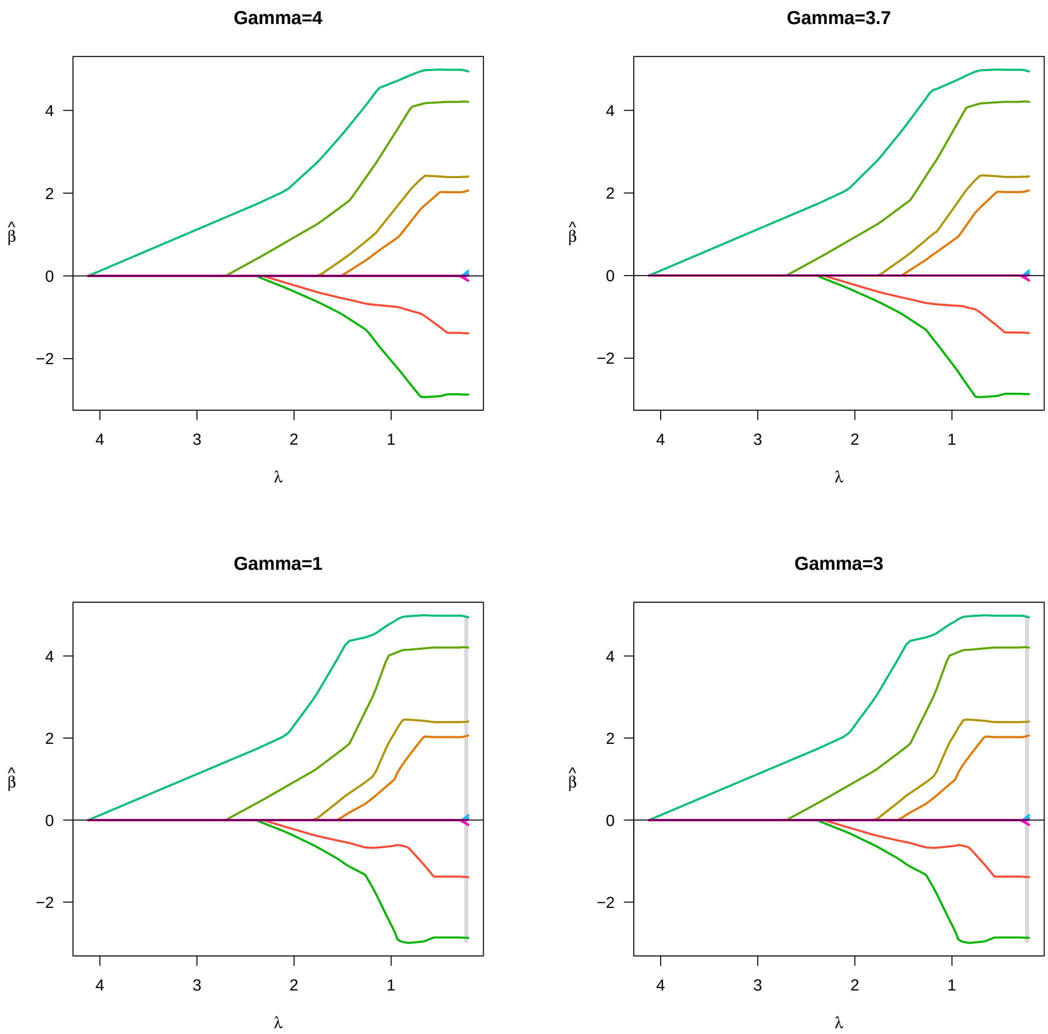
شکل ۹: نمودار اعتبارسنجی برای داده‌های شبیه‌سازی‌شده به روش شبکه ارتجاعی.



شکل ۱۰: نمودار برآورد ضرایب داده‌های شبیه‌سازی‌شده شبکه ارتجاعی.



شکل ۱۱: نمودار اعتبارسنجی داده‌های شبیه‌سازی‌شده برای پارامتر λ با روش SCAD.



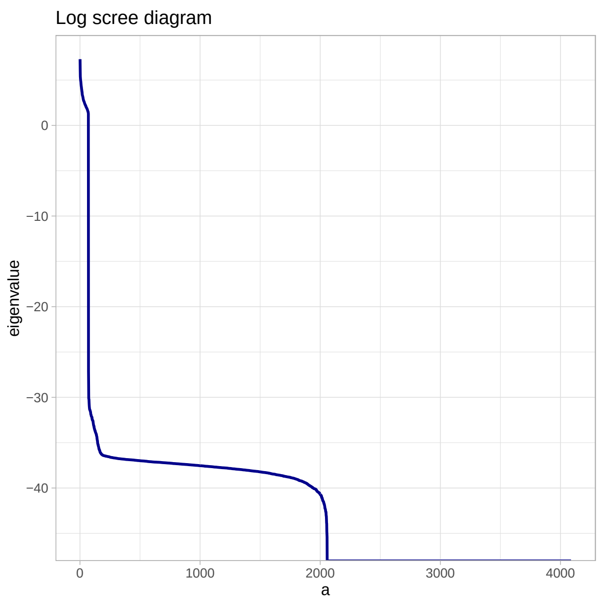
شکل ۱۲: نمودار برآورد ضرایب داده‌های شبیه‌سازی‌شده به ازای پارامترهای تاوان مختلف به روش SCAD.

جدول ۲: جدول نتایج روش‌های تاوانیده برای داده‌های شبیه‌سازی‌شده

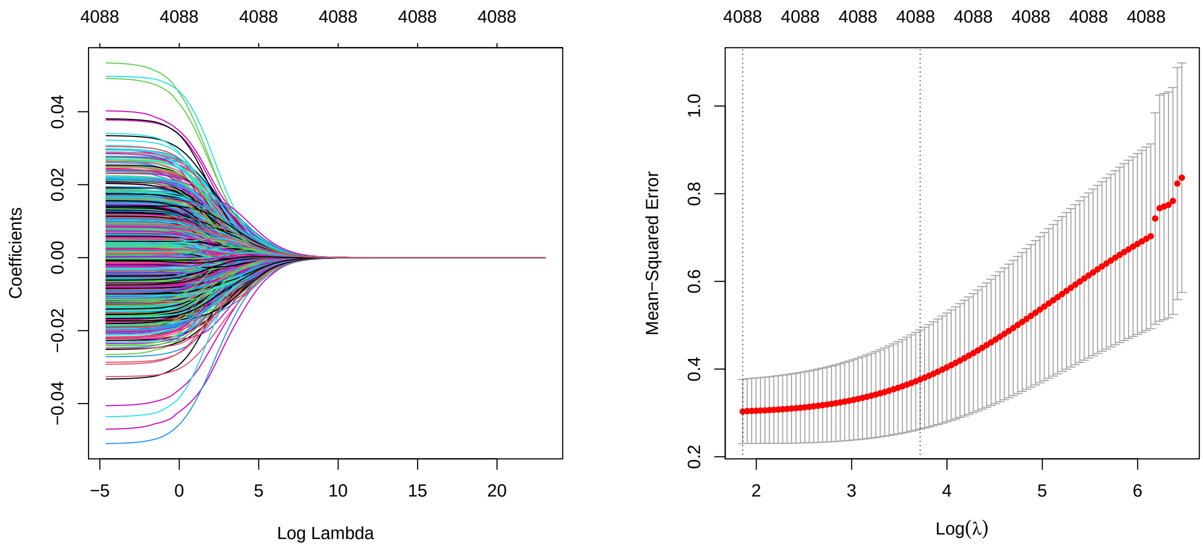
روش	مقدار بهینه پارامتر تاوان	تعداد ضرایب غیر صفر	مجموع مربعات خطا
ستیغی	۲۳/۱۰۱۳	۱۰۰	۱۹۳/۱۳۰۹
لاسو	۰/۱۰۹۷	۲۸	۲۲/۳۱۵۷
شبکه ارتجاعی	$\alpha = ۱, \lambda = ۰/۰۹۲۹۸$	۵۶	۵۳/۸۷۵۹
اسکد	$a = ۳/۷, \lambda = ۰/۵۰۶۷$	۶	۴۱/۹۱۴۰۵

بعضی از محصولات غذایی اضافه می‌شود و به‌عنوان یک مکمل رژیم غذایی در دسترس است. این ویتامین یکی از اجزای اساسی دو کوآنزیم اصلی، فلاوین مونونوکلئوتید (FMN) و فلاوین آدنین دینوکلئوتید (FAD) است. این کوآنزیم ها نقش عمده‌ای در تولید انرژی، عملکرد سلولی، رشد و نمو، و متابولیسم چربی‌ها، داروها و استروئیدها دارند. علاوه بر این، ریبوفلاوین به حفظ سطح طبیعی هموسیستئین، یک اسیدآمینه در خون کمک می کند. بیشترین درصد ریبوفلاوین پس از مصرف مواد گوشتی، سبزیجات، لبنیات و مقداری کمتر در مغزها و تخمه‌ها، حبوبات و سبوس غلات موجود توسط بدن دریافت می‌شود. بیشتر ریبوفلاوین توسط روده کوچک جذب و مقدار کمی در کبد، قلب و کلیه‌ها ذخیره‌شده و مابقی از طریق ادرار از بدن دفع خواهد شد. وضعیت ریبوفلاوین در افراد سالم به‌طورمعمول اندازه‌گیری نمی‌شود. اما در صورت بروز مشکل و تشخیص کمبود (یا وفور) این ویتامین از طریق بررسی گلبول‌های قرمز خون و یا بررسی میزان ریبوفلاوین دفع شده از طریق ادرار اندازه‌گیری خواهد شد. میزان میانگین ریبوفلاوین موردنیاز بدن روزانه بین ۱/۱ تا ۱/۳ میلی‌گرم برای بزرگ‌سالان و برای زنان باردار یا شیرده ۱/۶ می‌باشد. شایع‌ترین علت کمبود ریبوفلاوین در بدن رژیم غذایی نامناسب است. کمبود ریبوفلاوین هم‌چنین می‌تواند در افرادی که دچار نقص در فعالیت‌های کبدی هستند ایجاد شود، چراکه مانع از استفاده مناسب از ویتامین‌ها می‌شود. این داده‌ها در بسته نرم‌افزاری ''hdi'' نرم‌افزار R موجود است. در این مجموعه داده متغیر پاسخ لگاریتم نرخ تولید ریبوفلاوین است. این مجموعه داده دارای ۴۰۸۸ متغیر توضیحی بوده که هر کدام نشان‌دهنده لگاریتم سطح ژن‌ها است.

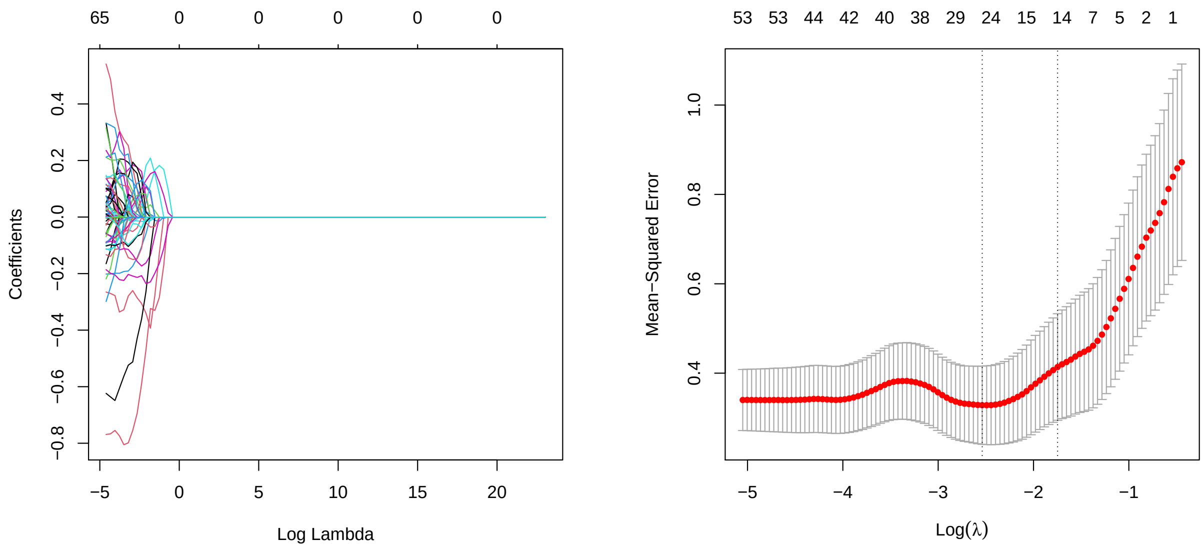
برای تحلیل این داده‌ها ابتدا از روش مؤلفه اصلی استفاده می کنیم. نمودار لگاریتم بازو حاصل در شکل ۱۳ قابل‌مشاهده است. در ادامه از روش‌های کمترین توان‌های دوم تاوانیده استفاده می کنیم. برای تعیین پارامتر جریمه از روش اعتبارسنجی متقابل ۱۰-سطحی با در نظر گرفتن ۷۰٪ داده‌ها به‌عنوان مجموعه آموزش و ۳۰٪ به‌عنوان مجموعه آزمون استفاده می‌شود. نمودارهای مربوط به اعتبارسنجی و برآورد ضرایب به ازای پارامترهای تاوانیده مختلف با روش ستیغی و لاسو مشهود است. در نمودار اعتبارسنجی روش ستیغی به ازای مقدار ۶/۴۰۹۵۳۶ مقدار MSE را کمینه کرده و در روش لاسو مقدار آن به ازای ۰/۰۷۹۰۱۹۷۹ کمینه شده است. در قسمت فوقانی نمودارهای سمت راست شکل‌های ۱۴ و ۱۵ تعداد ضرایب غیر صفر موجود به ازای مقادیر مختلف λ نمایش داده‌شده است که واضح است که با افزایش میزان λ تعداد ضرایب غیر صفر روش لاسو کاهش می‌یابد درحالی‌که در روش ستیغی برآورد ضرایب بسیار به صفر نزدیک می‌شوند ولی دقیقاً معادل با صفر نخواهند بود. نمودار اعتبارسنجی برای روش شبکه ارتجاعی در شکل ۱۶ نمایش داده‌شده است. نمودارهای برآورد ضرایب به روش شبکه ارتجاعی در شکل ۱۷ رسم شده است. نمودارهای برآورد ضرایب به روش اسکد به ازای دو مقدار $a = ۳,۳۷$ و بردار پارامتر دلخواه یکسان در شکل برای λ ۱۸ رسم شده است. نمودار اعتبارسنجی متقابل برای یافت مقدار بهینه پارامتر λ به ازای مقدار پیشنهادی فن و لی [۱۲] برای پارامتر a در شکل ۱۹ نشان داده‌شده است. در جدول ۳ نتایج روش‌های مورداستفاده گزارش‌شده است.



شکل ۱۳: لگاریتم نمودار بازو برای داده ریبوفلاوین.



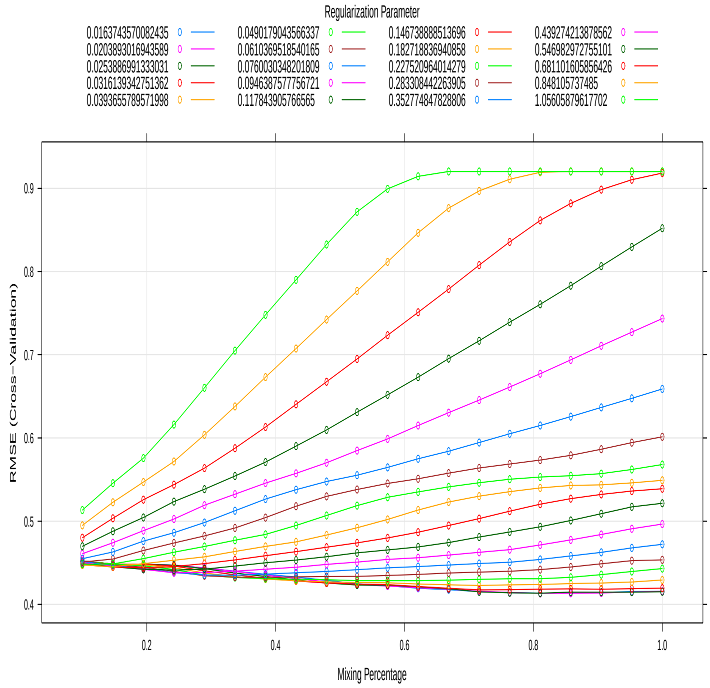
شکل ۱۴: نمودار اعتبارسنجی و برآورد ضرایب داده ریبو فلاوین به ازای پارامترهای تاوان مختلف به روش ستیغی.



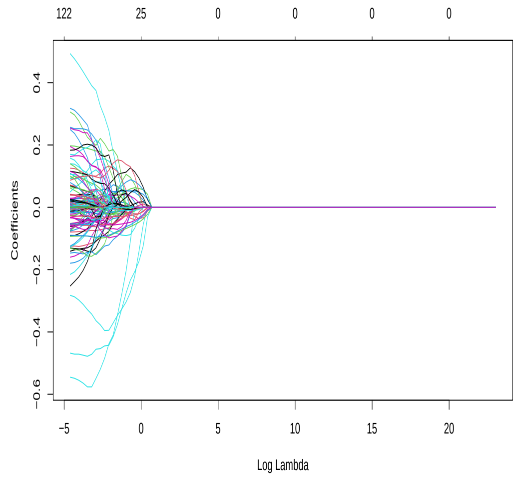
شکل ۱۵: نمودار اعتبارسنجی و برآورد ضرایب داده ریبوفلاوین به ازای پارامترهای تاوان مختلف به روش لاسو.

جدول ۳: جدول نتایج روش‌های تاوانیده برای داده ریبوفلاوین.

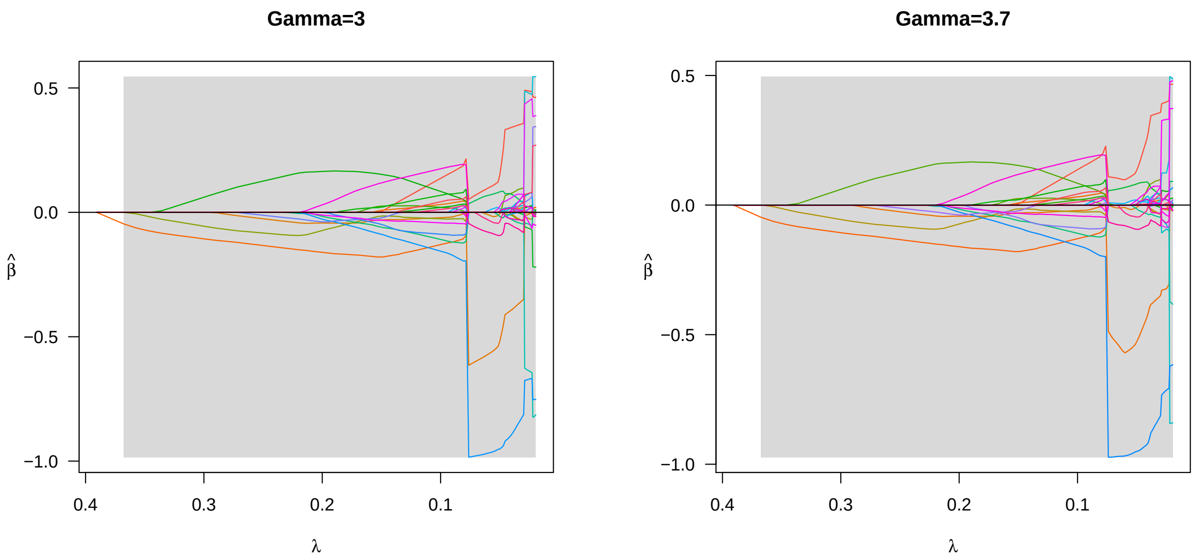
روش	مقدار بهینه پارامتر تاوان	تعداد ضرایب غیر صفر	مجموع مربعات خطا
ستیغی	۶/۴۰۹۵۳۶	۴۰۸۸	۸۴۶۱/۹۳۳
لاسو	۰/۰۷۹۰۱۹	۲۸	۱۰۲/۶۴۴
شبکه ارتجاعی	$\alpha =$ ۰/۸۱۰۵۲, $\lambda =$ ۰/۰۲۰۳۸	۴۳۷	۴۰۴۹/۶۰۵
اسکد	$a =$ ۳/۷, $\lambda =$ ۰/۰۵۷۱۵۱۱۳	۱۰	۲/۰۶۸۴۶۱



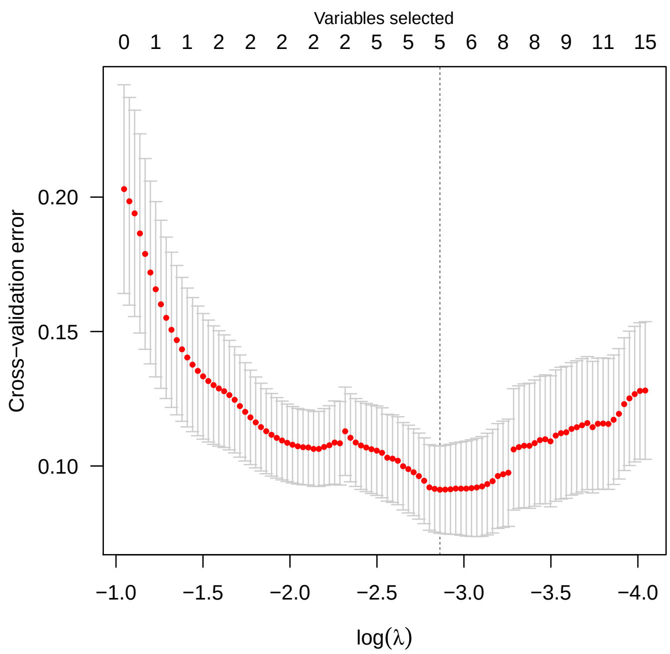
شکل ۱۶: نمودار اعتبارسنجی برای داده ریبوفلاوین به روش شبکه ارتجاعی.



شکل ۱۷: نمودار ضرایب برای داده ریبوفلاوین به روش شبکه ارتجاعی.



شکل ۱۸: نمودارهای روش اسکد برای داده ریبوفلاوین به ازای $\gamma = ۳$ و $\gamma = ۳/۷$.



شکل ۱۹: نمودار اعتبارسنجی برای یافت پارامتر λ برای داده ریبوفلاوین به روش اسکد.

۵ نتیجه‌گیری

مشکل همخطی و یا بزرگ‌تر بودن تعداد متغیرهای توضیحی نسبت به مشاهدات، جواب‌های روش کمترین توان‌های دوم را به طرز قابل‌توجهی منحرف یا ناممکن می‌کنند. برای حل این مشکل در مسائل مدل‌سازی رگرسیون تاکنون روش‌های متعددی پیشنهاد و بررسی‌شده است که در این مقاله به معرفی مهم‌ترین آن‌ها پرداخته شد (از جمله سایر روش‌ها می‌توان به [۵] و [۲۱] اشاره کرد). در حالت کلی نمی‌توان گفت که کدام روش بر سایر روش‌ها برتری دارد، زیرا هر روش طبق شرایط داده‌های جمع‌آوری‌شده بهتر عمل می‌کند. مثلاً روش لاسو نمی‌تواند در مواردی که بین داده‌ها هم خطی وجود دارد خوب عمل کند ولی روش شبکه ارتجاعی در این شرایط بهتر عمل می‌کند. از طرف دیگر روش رگرسیون ستیغی اگرچه می‌تواند مشکل وجود همخطی بین متغیرهای توضیحی را حل کنم، ولی روش مناسبی برای انتخاب متغیر نیست (برخلاف روش لاسو). به همین دلیل هدف مقایسه روش‌های مطرح‌شده در این تحقیق نبوده و بیشتر ایجاد یک دید جامع در خوانندگان در مواجهه با این‌گونه داده‌ها است. بنابراین هدف اصلی تنها معرفی و به‌کارگیری هم‌زمان روش‌ها در داده‌های واقعی و شبیه‌سازی‌شده می‌باشد. اگرچه به‌کارگیری هم‌زمان چندین روش بسیار راهگشا خواهد بود و دید جامعی نسبت به انواع مشاهدات در اختیار محققان و کاربران قرار می‌دهد.

تقدیر و تشکر

نویسندگان مقاله ضمن تشکر از اعضای محترم هیئت تحریریه مجله، از پیشنهادها و نظرات ارزشمند داوران و ویراستار محترم مقاله که موجب ارتقاء سطح آن گردید کمال تشکر و قدردانی را دارند.

مراجع

[۱] آرشی، م.، صادقی، ح. و طباطبایی م.، (۱۳۹۸)، استنباط آماری، انتشارات دانشگاه صنعتی شاهرود، شاهرود.

[۲] رستا، ر.، چینی پرداز،ر.و راسخی، ع.ا.، (۱۳۸۹)، معرفی روش کمترین توان‌های دوم تاوان‌داده در انتخاب متغیرهای توضیحی، ششمین همایش ملی آمار، دانشگاه پیام‌نور، اهواز.

[۳] جذن،س. و امینی، م.، (۱۳۹۶)، برآورداستوار نسبت به مشاهده‌های دورافتاده در رگرسیون خطی در حضور هم‌خطی چندگانه، مجله اندیشه آماری، **۴۴**، ۹۳-۱۱۰.

[۴] نوری جلیانی، ک.، نوری، س.، محمد، ک.، نیکنام، م.ح.، محمودی، م.، آندونیان، ل. و اکابری، آ.، (۱۳۹۰). آنالیز جنگل های تصادفی: یک روش آماری مدرن برای غربالگری در مطالعات با بعد بالا و کاربرد آن در یک مطالعه همبستگی ژنتیکی جمعیت-پایه، مجله دانشگاه علوم پزشکی خراسان شمالی، (ویژه نامه آمار زیستی واپیدمیولوژی)، **۳**، ۹۳-۱۰۱.

[5] Akdeniz, F and Roozbeh, M. (2019). Generalized difference-based weighted mixed almost unbiased ridge estimator in partially linear models, *Statistical Papers*, **60(5)**, 1717–1739.

[6] Bellman, R. (1961). *Adaptive control processes*. Princeton university press, London.

[7] Bertsimas, D and Parys, B. V. (2020). Sparse high-dimensional regression: exact scalable algorithms and phase transitions, *Biostatistics*, **21(2)**, 219-235.

[8] Bonddel, H.D. and Reich, B.J. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR, *journal of the interntional biometric society*. **64(1)**, 115-123.

[9] Breiman L. (1995). Better subset regression using the nonnegative garrote, *Technometrics*. **37(4)**, 373-384.

[10] Efron, B. and Hastie,T. (2017). *Computer age statistical inference*. Cambridge University Press, Cambridge.

[11] Everitt, B. and Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. Springer, New York Dordrecht, Heidelberg, London.

[12] Fan, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*. **96(456)**, 1348-1360.

[13] Fan, J. (1997). Comments on wavelets in statistics: a review by A. Antoniadis," *Journal of the Italian Statistical Society*. **6(2)**, 131-138.

[14] Frank, E. and Friedman, J. (1993). A statistical view of some chemo- metrics regression tools, *Technometrics*. **35(3)**, 109-148.

[15] Hastie, T. J. and Pregibon, (1992). *Generalized linear models*. Eberly College of Science, London.

[16] Hoerl, A. E. (1962). Application of ridge analysis to regression problems, *Chemical Engineering Progress*. **58(1)**, 54-59.

[17] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems, *Technometrics*. **12(1)**, 69-82.

[18] Jolliffe, I.T. (2002). *Principal component analysis*. Springer series in statistics, Aberdeen.

[19] Li, B. and Yu, Q. (2009). Robust and sparse bridge regression, *Statistics and its interface*, **2(4)**, 481-491.

[20] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*. **2(1)**, 559-572.

[21] Roozbeh, M. (2018). Optimal QR-based estimation in partially linear regression models with correlated errors using GCV criterion, *Computational Statistics & Data Analysis*. **117**, 45-61.

[22] Roozbeh, M., Babaie-Kafaki, S. and Naeimi Sadigh, A. (2018). A heuristic approach to combat multicollinearity in least trimmed squares regression analysis, *Mathematical modelling*. **57(2)**, 105-120.

[23] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*. **58(1)**, 267-288.

[24] Walker, D. A. and Smith, T. J. (2020). Logistic regression under sparse data conditions, *Journal of Modern Applied Statistical Methods*, **18(2)**, 33-72.

[25] Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science and Business Media, New York.

[26] Watkins, D.S. (2002) *Fundamentals of matrix coputations*. John Wiley and Sons, New York.

[27] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B*. **67(2)**, 301-320.